Towards a Linguist's Workbench Supporting eScience Methods

A. Dimitriadis, M. Kemps-Snijders, P. Wittenburg, M. Everaert, S. Levinson
University of Utrecht, MPI for Psycholinguistics
alexis.dimitriadis@let.uu.nl, marc.kemps-snijders@mpi.nl

Abstract

The domain of language resources is fragmented in many dimensions. Institutional fragmentation is currently being addressed by Grid projects, which will allow access to resources across institutional boundaries. While technical encoding and structural/format differences constitute significant challenges,, this paper focuses on the problem of the terminological differences encountered when researchers access resources from different projects and creators. We outline two projects that employ a bottom-up approach, and discuss potential extensions towards an eventual Service Oriented Architecture that will bring together all the different components required to overcome the various fragmentation boundaries and open the road to an eHumanities environment.

1. Introduction

According to John Taylor,[1] eScience is "about global collaboration in key areas of science, and the next generation of infrastructure that will enable it". This means a new form of aggregation (1) in the way collaboration is carried out, since science is based on an open exchange of competing ideas and extensive scholarly interaction, (2) in the way existing boundaries in accessing a common domain of resources are overcome and (3) in the way new opportunities for cross-discipline fertilization are offered. The realization of the eScience vision will only be brought about through continuous innovation in Information Technology; in particular, through the increasing power of electronic networks, Internet-facilitated access to increasingly powerful distributed high performance computers, access to large distributed repositories of useful resources and knowledge, and increasingly smarter Semantic Web technologies. This scenario will be true for the typical "large scale problems", such as solving the questions in understanding human genetics, where all available resources have to be used in parallel to get new insights, as well as for the typical "small scale problems" where a researcher is looking for information on a comparatively small, newly posed question, nevertheless looking at different electronic resources housed at different locations.

The current infrastructure by itself, consisting mainly of powerful networks and Internet services such as email and the World-Wide-Web, will not be sufficient for all the expectations associated with the terms "eScience" to become reality. New types of infrastructures are needed, such as those indicated by the term "Grid[2]". There is no doubt that the natural sciences are the driving force in defining the key elements of such new infrastructures. "Computing Grids[3]" were in the focus of the integration work, allowing theoretical physicists and chemists to tackle Grand Challenges where new breakthroughs require the interconnection of high performance computers for some length of time, and the exchange of large amounts of data between processors operating at distinct locations.

From this perspective, data exchange and integration problems in natural science revolve mainly around the sheer mass of data and its technical and syntactical encoding. In the humanities, the major obstacle to data interoperability is syntactic and semantic heterogeneity. Roughly speaking, it is the differences in terminology that make it so difficult to cross the boundaries and create a joint domain of language resources that can be utilized seamlessly.[4] In this paper we present the solutions that were found in two different areas in linguistics, typological databases and corpus linguistics, to create integrated views of heterogeneous data. We further sketch how these two areas could be integrated with the help of a Service Oriented Architecture, which will open the possibilities of new forms of collaboration and cross-fertilization. The solutions found and the suggested integration will be a step towards an eHumanities infrastructure.

---

[1] UK eScience:
http://www.rcuk.ac.uk/escience/documents/report_coreproggrid.pdf

[2] Grid: http://www.gridforum.org/
[3] Grid Computing: http://www.grid.org/about/gc/
[4] This is one of several reasons suggesting a pre-paradigmatic phase of theorizing for these disciplines.

2. The Domain of Language Resources

Modern digital technology revolutionized the way language resources were created, collected and stored. Today it is comparatively easy to carry out computer-aided work in linguistics, and there is a wide number of options to carry out this work. With respect to their structure, language resources typically cover numerous possibilities:

- Databases with highly structured information are frequently used for lexica, terminologies and typology. Here we can speak of data that is compliant to a highly constrained entity-relationship model. The semantic content of such resources, however, is often less well-defined than the storage model.
- Much data is available in form of annotated multimedia corpora. While the annotation structures can become rather complex, the data is basically sequentially organized and is only partly constrained by controlled vocabularies. Underlying these annotations are multimedia recordings covering different streams of data such as audio, video, eye tracking data etc. So for the annotation data we can speak of semi-structured data.
- Many language resources cover large texts that are only coarsely tagged. Also here we can speak of semi-structured data, although the textual fragments are unstructured.
- Finally, we have resources such as grammars, descriptions of the phonetic system, etc., that are completely unstructured texts.

For the more structured data a wide range of different container types has been used, ranging from relational database management systems to document editors such as WORD, where structure is frequently indicated by font type etc. In addition, every research group created its own schema, tagging or encoding system, since there were no widely used standards or best practice guidelines. In addition, in many cases the tags used were/are not documented, i.e., only the researchers who created a resource can interpret it. For all efforts to cross the boundaries between individual resources or coherent collections that have emerged from projects such as the Dutch Spoken Corpus[5], the researcher has to function as an interface. This is particularly true for language resources that were created not for the purpose of being shared and re-used, but as the byproduct of individual research activities. Thus,

no automatic processes such as, for example, searching or making statistics can be carried out across such boundaries. This, and the need to know about the idiosyncratic details of each user interface, had made it practically impossible for researchers to make use of the rich domain of language resources in their research work.

3. Crossing the Boundaries

Given this situation, a number of initiatives were started to overcome the various boundaries. Grid projects such as DAM-LR[6] are working to overcome the institutional boundaries, i.e., to allow users to navigate in integrated metadata catalogues, to support users with a single identity accepted at all participating archives, and to create a coherent domain of unique and persistent resource identifiers, all based on a network of trusted servers and services. Other projects were started to develop tagging and metadata standards, such as TEI[7], IMDI[8] and Dublin Core[9]/ OLAC[10], to design generic models such as LMF (Lexicon Markup Framework)[11] and to work out recommendations for linguistic encoding with the help of Data Category Registries (DCR), such as within ISO TC37/SC4[12]. All these projects have as their goal to provide general frameworks that can overcome the fragmentation of language resources. However, it will take some time to make researchers aware of these possibilities. Furthermore, we will have to deal with so-called legacy data, i.e., data that do not conform to established standards. One might argue that it is necessary to transform all legacy material into more formalized representations that can be accessed and managed automatically. However, we understand that these resources are still heavily used and continuously enriched, that "legacy" data continues to be created, and that each formalization is associated with a loss of information, i.e., we have to ensure that the original versions will be maintained.

In addition to the above initiatives, two projects were started in the Netherlands that tackle the semantic interoperability issues with bottom-up, data-driven approaches. Their primary goal is to create an integrated domain of language resources that unifies a variety of (semi-)structured legacy resources such as typological databases, lexica and annotated media resources. Although the difficulties of overcoming the

[5] CGN:
http://ww2.tst.inl.nl/index.php?option=com_content&task=view&id=244&Itemid=261

[6] DAM-LR: http://www.mpi.nl/dam-lr
[7] Actually, TEI has existed for quite a while, but only recently received the attention that is necessary to achieve a higher degree of unification at the encoding level. http://www.tei-c.org/
[8] IMDI: http://www.mpi.nl/IMDI
[9] Dublin Core: http://www.dublincore.org/
[10] OLAC: http://www.language-archives.org/
[11] LMF: http://estime.spim.jussieu.fr/~pz/lrec2006/Francopoulo.pdf
[12] ISO TC37/SC4: http://www.tc37sc4.org/

format and structural problems are considerable, this paper focuses on the semantic aspects, i.e., the strategies used to overcome the differences in linguistic encoding, tagging, and analysis.

At the MPI for Psycholinguistics, a language resource archive is being maintained that currently contains about 250.000 objects[13], mostly annotated media resources, but also lexica. These resources have been created by many different researchers who worked in many different projects. The archive is open for deposits by researchers at MPI or at other institutions without archiving facilities. Due to ongoing projects this archive is continuously being extended in various ways, partly by enriching or correcting the existing linguistic encoding, partly by adding new collections. As indicated above, all researchers and projects are independent in their choice of how linguistic phenomena are encoded. Therefore, we cannot speak of a closed domain of semantic concepts but of an open domain where new concepts are introduced, where names that already exist in the archive are used to express different meanings, and where different names are used to identify the same or rather similar meanings. The question that was tackled is which semantic interoperability mechanisms have to be available for a researcher when he or she wants to carry out, for example, searches across several of these collections, knowing that the encoding and content of the resources will vary continuously.

The goal of the TDS project[14] is to provide integrated access, through a web-based service, to a virtually integrated domain of typological databases. Each such database contains a very large number of data fields, typically several hundred, about a large number of languages (again in the hundreds). Also here we can state that the typological databases were created independently of each other, with a focus on different aspects of languages and with different intentions in mind. Similar issues of semantic differences and similarities arise as in the MPI case. However, in this case we are dealing with a relatively small number of resources (databases), each semantically complex and containing relatively high-value information. There are about a dozen databases in the initial phase of the project, and the eventual size of the archive will be in the dozens rather than thousands of databases. Hence the focus was on unifying the semantics and encoding of a particular (but progressively extended) set of data-

bases, i.e., the semantic scope of all concepts that are used within an initial set of typological databases could be carefully studied and analyzed.

Both initiatives focused at the first instance on web-applications that allow users to operate in the created integrated domain via web-based interfaces.

4. Solution for Language Resources

4.1 Semantic Interoperability

The interoperability task has to be driven by the current selection of resources, since there is no way to carry out an analysis of the concepts used in all resources: The archive has too many resources, and it is continuously being extended. The resources covered in the MPI archive are logically organized with the help of linked metadata descriptions. When ingesting resources, canonical trees are defined by the researcher and represent the organization of the data. In the figure below a sub-tree is shown that points to the CGN project (Dutch Spoken Corpus) The CGN is a national project with clear guidelines of how to select the tags and how to encode linguistic phenomena at levels such as phonetics, orthography, morphology, syntax and prosody. It has its own internal structure supporting easy navigation etc, but we can assume that we can associate a "Concept Profile" with the top level node which includes all concept definitions, including value ranges where possible. These definitions are valid for all resources in the appropriate sub-tree.
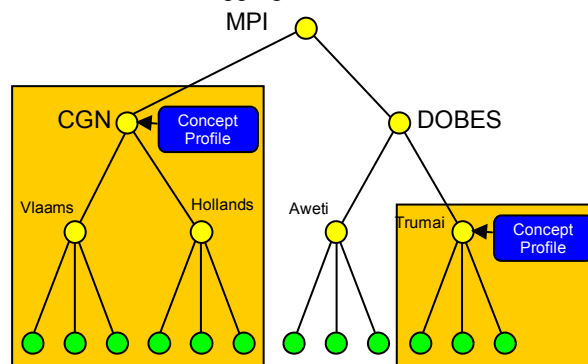
Another sub-tree points to the DOBES project[15] (Documentation of Endangered Languages), covering documentation from a number of languages that are endangered (such as Aweti and Trumai) and will probably become extinct in a few years. Since these languages are mostly very different from each other in many ways and since the documentation teams work in isolated circumstances at rather different locations, it was not feasible to agree on a set of well-defined con-

cepts. Therefore, we can expect that each team has defined its own set of concepts to be used for encoding linguistic phenomena. However, we can also assume that for many deposited collections in the archive there is no well-defined set of concepts. This situation is indicated in the figure, where the Trumai node is associated with a "concept profile" and the Aweti node is not, i.e., in the latter case it is left to the user to find out, by studying the content, which concepts are used and what they mean.

In some cases, then, there are excellent formal, machine readable definitions of the concepts used in a whole sub-tree of resources, while in other cases there are no formal definitions. It should be added that the idea of machine readable concept profiles is new to most linguists, i.e., in general there were no such profiles, but informal descriptions are now gradually being turned into formal descriptions.

Suppose now that a user selects two resources from domains that have concept profiles: A crawler automatically moves up in the canonical tree to find appropriate Concept Profiles (CP), and users now need a framework that allows them to easily create Personal Concept Registries (PCR) by selecting concepts from the different CPs, relating them to each other, and integrating them into their own concept work space. In those cases where no CP can be found by the crawler, the users should get some information about the names found in the resources. Again these will be included in the PCR, but users have to find for themselves what their exact meaning is.

The next logical step is for users to record discovered "personal" semantic relations between selected concepts. The framework should support this in a smart way, so that editable Personal Relation Registries (PRR) are created, which should have the capability of being made persistent in the same way as the PCRs. These relations can be used for example by a search engine to operate across collection boundaries. To make this work, the PRRs have to include a relation type specification and the PCRs a unique identification of the resource domain from which the concept was extracted.

It is obvious that the creation of PCRs and PRRs can become a very time consuming activity; hence, all options to increase efficiency should be exploited. For some concepts extracted from CPs or found in resources it will be the case that there are excellent definitions in central Data Category Registries (DCR) such as from ISO TC37/SC4. In these cases it would be sufficient to insert a link to such an entry. If two extracted concepts point to the same DCR entry, for example, an equality relationship can be assumed and used during operations. The framework should indicate such equalities. Over time there may also be an institutional

DCR. Here the same holds: PCR entries pointing to the same entry in another "central" DCR implicitly mean "equality".

Another important aspect is the possibility to edit, store and exchange such PCRs and PRRs with the help of the framework. The canonical archive structure at the MPI is an excellent means to store and share such knowledge components.

A first version of an Ontology Editor was built that includes the most essential features described above. The screenshot below shows the concepts included in two CPs (two panes at the right) and a user-made list of concepts (left pane). On the top of each pane one can see those concepts that are selected and that are related with other concepts.
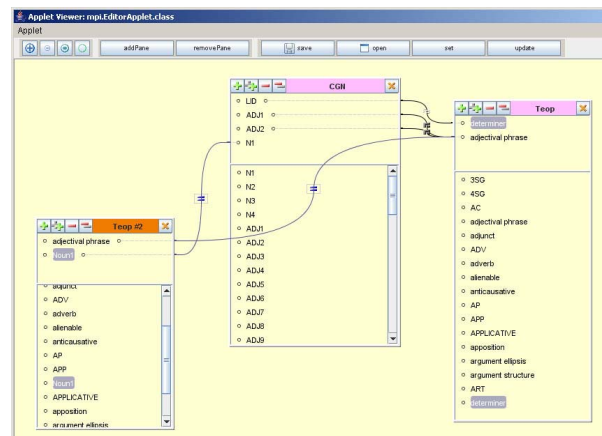
When saving, two XML files are generated: the PCR will contain all concepts selected, including the information from where they were extracted, etc, and the PRR will contain the relations. The editor indicates the relation types graphically. Our experience with linguistic data has shown that four relation types are sufficient at this stage: "is_equivalent_to", "is_generic-_to", "is_specific_to" and "relates_to". "is_generic_to" indicates broader generic term relations such as between "verbs" and "transitive verbs", "is_specific_to" indicates the inverse relation and "relates_to" indicates some kind of similarity between the concepts that cannot be expressed in more detail. The editor allows users to store and re-use PCRs and PRRs; however, registering and storing them in an archive sub-tree for sharing purposes is yet to be implemented.

4.2 Web Applications

The MPI currently offers a number of web applications/ services for accessing the archive:

- metadata applications (native XML IMDI

Browser[16], XSLT transformation to HTML, IMDI Search) allow users to browse and search (structured and unstructured) for suitable resources in the archive by making use of the IMDI schema and vocabularies;

- a service provides OAI PMH[17] compliant metadata records;
- ANNEX[18] provides services for searching (structured and unstructured) through all structured annotations in the archive and visualizing the annotations fragments including the corresponding multimedia segments.
- LEXUS[19] provides equivalent services for structured lexica with multimedia extensions

The searching components in ANNEX and LEXUS will include the knowledge components mentioned above, since they allow one to operate on resources coming from different contributions. Currently, a first simple interoperability version has been implemented to test the potential and interest.

5. Solution for Typological Resources

While the MPI archive is designed to accommodate a large number of independently collected and marked-up corpora, the Typological Database System (TDS) focuses on optimal integration of a relatively small, selected collection of typological databases. A typological database typically contains highly distilled, general information about a large number of languages. A database in the TDS system might contain up to several hundred variables such as "this language has subject-verb agreement in the present tense", with values for between one and several hundred languages. Several databases also contain glossed sentences from numerous languages.

The TDS is an ongoing research project, whose aim is to develop a web-based service for unified querying of a collection of such independently created typological databases. The prototype server currently contains information on circa 1,000 languages from six integrated databases. Its component databases contain data on a range of linguistic topics including agreement, parts of speech, word order, stress placement and predication phenomena. Other databases also contain primary linguistic data in the form of lexicons and glossed sentences. To facilitate data integration and management, the TDS relies on an ontology of linguis-

tic Concepts that the project has developed. To compensate for the differences between the databases, the TDS Ontology (TDSO) provides a non-prescriptive, or "inclusive" framework of linguistic concepts and terms, into which the particular perspective of each component database can be integrated. Explicit links between the unified data and ontology concepts facilitate searching through the integrated database fields.

The goals of the TDS system are (a) to provide an interface that will help users *find* relevant data, and (b) to allow users to *interpret* the data they are presented with. The TDS Ontology is utilized in both tasks: Searching for data relevant to a topic is mediated by links between database fields and values, and Concepts in the ontology. Interpretation of the data is supported by the Concept documentation, which may be presented alongside a database field's own documentation. As data may be presented out of their original context, in all cases the interface must provide the provenance of the data along with database-specific description; this allows users to properly evaluate the information presented.

Searching is a two-step process. First, the user discovers fields relevant to the topic being researched, by using one of the searching or browsing options provided by the TDS interface. Selected fields are accumulated, forming a pre-query. In the second step, the user refines this pre-query and executes it.

5.1 Overview of the system

An overview of the TDS architecture is shown in the figure below. At the right are the system's ontologies, which guide the management of the data. The system relies on a *hybrid,* or *two-level,* ontology design: At the top is the global TDS Ontology of Linguistic Concepts, which is not prescriptive but provides a global frame of reference. It is extended by the database-specific *local ontologies,* which include the idiosyncratic definitions applicable to each database; the local ontologies are an integral part of the local database schemas ("DTL specifications"), which specify the mapping of database contents to the TDS space.

The system operates by periodically importing the contents and metadata of the component databases. These are restructured, merged and transformed into a single hierarchical data structure, *i.e.,* a tree. The user interface interacts with this tree, and with its associated specification schemas, extracting information through queries. To create the global tree, each database is first addressed on its own, *i.e.,* a collection of trees is built with each tree containing the data of one component database. The schema for each of these trees is described using a special-purpose language we have developed, called the *Data Transformation Language*
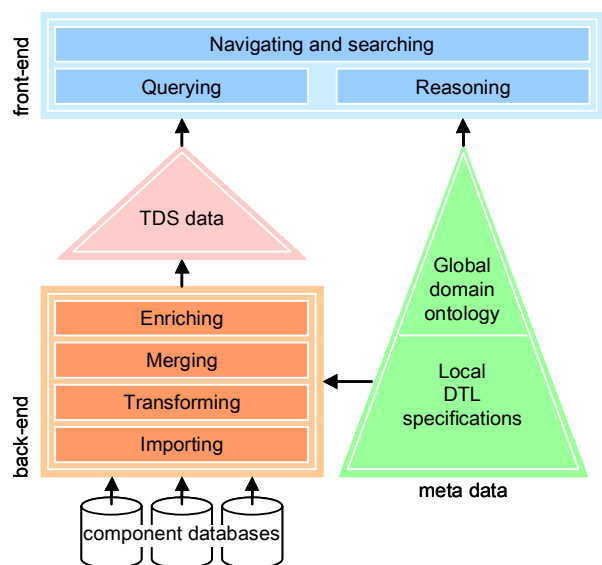
---

[16] IMDI-Browser: http://www.mpi.nl/imdi/tools/
[17] OAI PMH:
http://www.openarchives.org/OAI/openarchivesprotocol.html
[18] ANNEX: http://www.mpi.nl/annex
[19] LEXUS: http://www.mpi.nl/lexus

front-end

Navigating and searching

Querying | Reasoning

TDS data

Enriching

Merging

Transforming

Importing

back-end

Global domain ontology

Local DTL specifications

meta data

component databases

(DTL). A DTL specification describes Notions and their relationships, and the nodes in the tree are thus instantiations of these Notions. The semantic content of the tree schemas constitutes the local ontologies.

The internal organization of the DTL specification is designed to parallel, as far as possible, relevant parts of the global linguistic ontology. Notions can be shared by the various DTL specifications, and their instantiations can be shared as well, by means of keys where necessary. This allows the collection of trees to be merged into a single one. In addition to the data actually in the component databases, the tree may also be enriched with *derived Notions,* which are field Notions computed on data from one or more component databases. The result of these steps is a single hierarchy containing all the data from the integrated databases, shown in the diagram as the *TDS data*.

The query system interacts with both the TDSO and the DTL specifications, allowing a user to discover database fields of interest and thus to formulate and execute a query. Selected fields are collected into a pre-query that can be further refined and carried out.

The implementation of the TDS relies on a range of technologies. The system uses a plug-in model to import data from diverse database sources, including Microsoft Access, XML data sources, MySQL, and comma-separated values (csv). The global ontology is maintained in the OWL ontology format, managed interactively with the Protege ontology editor, and enriched and validated through special-purpose batch tools. The local ontologies, on the other hand, are an integral part of the mappings between databases and the TDSO, and are defined in the special-purpose Data Transformation Language (DTL) developed by the project.

5.2 Differences addressed

We close this section with a discussion of the different kinds of variation that a linguistic integration system must address, and the strategy followed by the TDS:

Different linguistic data types. While the MPI archive consists primarily of annotated texts of various lengths, so-called "analytical" typological databases consist of high-level logical variables describing each language as a whole; for example, "non-finite verbs take genitive subjects". Other TDS databases contain example sentences with detailed annotations ("sentence databases"), or a combination of the two. The TDS is designed to integrate different types of linguistic data so that, for example, a single query can search both examples and logical variables for relevant information.

Different theoretical commitments. The information in the various databases reflects the analytical and theoretical commitments of its creators. While purely notational differences can be bridged with the help of detailed machine-readable metadata, conversions between theories cannot be automated with any reliability, and are often impossible in principle without loss of information. Therefore it can be useful for users to view information that is not expressed in terms of their theoretical framework. For example, information about the properties of "subjects" can be useful even to linguists who do not believe that *subject* is a well-founded notion. Such information will allow a knowledgeable user to recognize the descriptive content of a statement about language, and to gain useful information from it even if it does not exactly match one's own theoretical orientation.

The proper handling of such differences is at the center of both projects. Rather than try to convert this diverse information into a common framework, the TDS places a high priority on preserving and presenting to the user the framework of database-specific assumptions required to properly interpret the data in a component database.

In addition, there are the usual **non-semantic sources of differences,** which (as already announced) are not the focus of this paper. Typological databases focus on different phenomena, rely on different database software or operating systems, make different design choices, and encode the same values in different ways. The general approach of the TDS is to compensate, wherever possible, for purely notational variation, including design decisions and software platforms. Differences in theoretical orientation, however, must be preserved and presented to the end-user.

5.3. MPI vs. TDS

It can be seen that the two collections we discuss must deal with heterogeneity of quite different sorts. The MPI archive consists in particular of a very large number of annotated texts and highly structured lexica, each of which can be considered to be drawn from a single language, and to have its own notational conventions. The TDS is designed to accommodate a small number of databases, counted in the dozens rather than the hundreds or thousands; but each database contains information about hundreds of languages, and has more complex and (therefore) less predictable structure than the corpora and lexica archived at MPI. Nevertheless, in both cases it makes sense to treat the set of semantic concepts, and their properties, as essentially open and potentially irreconcilable; rather than try to eliminate the differences between them, an integration system must anticipate them and provide a way to manage them intelligently.

The systems necessarily differ in their approach to the integration process. While the MPI archive primarily relies on a lightweight process initiated by the corpus creator, the TDS utilizes detailed metadata and semantic categories, whose creation requires considerable knowledge of the system. While some tools are being developed to simplify this process, creation of the metadata requires collaboration of a project expert with the creators of each component database.

6. New Options and Opportunities

As we have seen, both of the projects discussed have relied mainly on data-driven bottom-up methods to achieve semantic interoperability. Moreover, both systems rely on custom semantic descriptions as a fallback when the data does not match an existing schema; the process could achieve a great gain in efficiency if it could rely on existing data category registries or ontologies for a large proportion of the incoming data. This would also make possible a number of new functions for end users: cross-corpus or cross database searches, statistics on larger collections, much easier comparison of linguistic encodings, etc. A standardized Service Oriented Architecture would allow many more such resources to be pooled together, breaking more of the existing boundaries and allowing a larger group of interested researchers to access the data through web applications. In the following we focus on a few examples, which must be seen as merely indicative of the directions to be pursued.

6.1. Across linguistic data types

The web platform allows us to support operations that cross the boundaries of linguistic data types. Yet, there is no web-based framework that allows researchers to jump from annotations to lexica and to typology databases[20]. Some steps have already been taken to connect lexical and annotation information via web-based services in the MPI archive, and to integrate analytical and text-based data in the TDS. It is easy to imagine many extensions of this approach:

1) Searching in multiple Typological Databases, e.g., for a certain morphosyntactic pattern. A single search operation should be able to look for analytical information, such as "this language has case marking", and also to look for known case markers (e.g., "accusative") in the morphosyntactic annotations of included sentences. The results might be presented as a list of languages that may have the searched-for property; the researcher can then look in detail at the annotations, and when appropriate even listen to underlying speech recordings, to check whether the typological characterization is correct. The researcher may then want to annotate the archive with references to typical examples, for the benefit of later users.

2) Making the lexicon the central anchor point for the documentation of a language.[21] To document the meaning of words, the lexicon will link to various media fragments demonstrating the cultural background of the word's meaning. A researcher can select a word in the lexicon, search for all occurrences of this word in all texts for the given language, evaluate the hits and add references to appropriate fragments.

3) A lexicon includes morphosyntactic encodings and can be used for semi-automatic annotation, i.e., whenever a new word is found in the transcription the appropriate morphosyntactic information is added.

6.2. Across Archives

The interoperability solution pursued at the MPI can easily be extended to a situation where different archives are connected via Grid technologies. Resources selected from other archives can be treated like just another collection within the same archive. Such external archives may or may not provide Concept Profiles; if not, this would decrease the efficiency of the system but not prevent access. A crucial component is the creation of an organizationally integrated domain, the task of Grid components as implemented in the DAM-LR (Distributed Access Management for

---

[20] We purposely exclude some toy implementations in very complex database designs, where people tried to incorporate all types of linguistic information, ignoring the fact that such solutions are not feasible from a software engineering point of view.
[21] The MPI is currently pursuing two such projects on Multimedia Encyclopedic Lexica, i.e., lexica with references to (annotated) media segments and layers of semantic relations.

Language Resources) project[22].

6.3. Collaboration

Web-based architectures provide a good basis for collaboration mechanisms, i.e., to support collaborative work on complex linguistic data types. MPI's LEXUS application already has collaborative mechanisms to work jointly on the creation of lexica. The greatest problems are of organizational nature: How to ensure that the lexicon remains in a consistent state when remotely operating colleagues manipulate either structure or content. The problems that may occur have been studied in the realm of large transaction systems, and solutions have been suggested. In addition to online transaction systems we are faced with the requirement that researchers will work on an offline copy for a while, and then return a new version to share it with the colleagues. Smart and interactive merging facilities are required to create new and accepted master copies.

6.4. Commentary

Web-based architectures are also suitable for implementing commentary mechanisms, i.e., allowing (authorized) users to add comments or annotations to any content or to draw typed relations between content fragments[23]. Also in this case the major problems are of an organizational nature. The MPI is taking the first steps toward a flexible commentary framework.

7. Conclusions

The interoperability and distributed editing scenarios described above assume sophisticated tools that can manipulate a common substratum of diverse language resources, through an open interoperability layer. Such tools must be built on top of a service-oriented architecture, that will make the above scenarios possible and allow developers and researchers to flexibly combine resources and operations in new ways.

Assuming that there are services that allow searching for specific acoustic/phonetic patterns in a speech recording, a typologist studying the sound repertoire of a certain language could look for an acoustic/phonetic pattern as described in a typological database, turn it into a suitable representation and then search for this pattern by integrating the appropriate services into a single framework, i.e., by starting appropriate methods and providing the necessary data. A wide range of similar options can be thought of. The basis of a boost

for the research work is the open accessibility of services of all kinds, via well-described interface specifications and mechanisms to find them.

We see a new type of linguistic workbench emerging, where the linguist's screen is the meeting point of all sorts of useful information, generated by a rich framework of services aggregated to new types of applications. These applications run on top of a Grid of repositories and service centers that helps cross the current institutional boundaries, i.e., for the researcher these institutional boundaries will become transparent. The provided services, specified by standardized interfaces, are building blocks that can be easily combined by application builders to overcome the remaining interoperability problems. In particular, a framework will be made available that easily allows users to create, manipulate, store and share knowledge components that can be integrated into operations such as advanced searching. In doing so, other applications can be envisaged that will bring the researchers unprecedented power over a large virtually integrated resource base.

However, we also foresee that much education and training effort will be needed to inform scholars and students about the emerging opportunities and to engage them in the new paradigm. Only broad acceptance by software developers and end-users will make it possible to achieve the huge task of integrating the existing resources and algorithms. Fully compliant with the vision indicated by John Taylor, cited in the introduction, we have explained that for us the term "global collaboration" includes the notion of being able to access a large domain of virtually integrated resources and functions and that the "next generation of an infrastructure" extends to a Service Oriented Architecture built on top of a Grid infrastructure.

---

[22] DAM-LR: http://www.mpi.nl/dam-lr
[23] We use "commentary" as a cover term, since relations are just a special form of commentary.