

Aspects of an XML-Based Phraseology Database Application

Denis Helic¹ and Peter Ďurčo²

¹ University of Technology Graz
Institute for Information Systems and Computer Media
`dhelic@iicm.edu`

² University of St. Cyril and Methodius Trnava
Department of German Language and Literature
`durco@vronk.net`

Abstract. In this paper we discuss technical aspects of a phraseology database application that is being developed in the scope of the Ephras project. Thereby we give a special attention to the data modeling perspective of such an application. We argue that the phraseology data is simply a particular kind of semi-structured data. Therefore, this data should be represented and managed by technologies that are specifically targeted at management of semi-structured data. Currently, the most prominent of such technologies is eXtensible Markup Language technology. Thus, in the remainder of the paper we discuss different implications of this technology on the application architecture and its implementation.

1 Introduction

The Ephras project is a project funded by the European Commission under Socrates/Lingua2 programme. The goal of the project is to develop a computer supported phraseology learning material for four European languages - German, Slovak, Slovenian and Hungarian language. The project aims to eliminate the lack of such phraseology learning material, as well as to meet the demands for multilingual learning material in the enlarged European Union. The Ephras learning material will be composed of a searchable database of 4x1000 phraseology data items in four languages (i.e., 1000 data items in each of the languages) accompanied with 150 interactive tests to selected phrases in four languages.

In this paper we concentrate on the first component of the Ephras learning material - the Ephras phraseology database application - by discussing its requirements and features, as well as a number of important technical issues related to that component. The most important requirements and features of this application can be summarized as following.

Firstly, the source language is German with 1000 phraseology data items, where each of these data items is a single phrase in German with one or more meanings. Additionally, each German phrase is involved in a so-called equivalence relation with data items from other three languages (target languages). The

equivalence relation expresses rather complex phraseology relationships between different data items. It can represent one of the following:

- A one-to-one relation between a single-meaning German phrase and a single-meaning phrase from any of the target languages.
- A one-to-many relation between a single-meaning German phrase and a number of different single-meaning phrases from any of the target languages.
- A one-to-one relation between a multiple-meaning German phrase and a multiple-meaning phrase from any of the target languages.
- A one-to-many relation between a multiple-meaning German phrase and a number of different single-meaning or multiple-meaning phrases from any of the target languages.
- Sometimes there is no direct phraseology equivalent for a German phrase in the target languages. The equivalents in that case are non-phraseology data items, i.e., a single word or a free collocation. Thus, the equivalence relation in such a case is a one-to-one relation between the German phrase and its corresponding free collocation.

Thus, the equivalence relation is 2-dimensional, where in the first dimension we have a one-to-one or a one-to-many relation between a German data item and the corresponding data items in the target languages. Orthogonal to that relation there is a relation between meanings of different data items, which can be either single-meaning or multiple-meaning data items, i.e., this relation is a typical many-to-many relation.

Further, in addition to the primary direction of the equivalence relation (i.e., the direction from German to other three languages) the secondary direction of this relation can be established as well (see Fig. 1). In some special cases (e.g., when only one-to-one relations are present) it is possible to infer the equivalence between data items from the target languages. For example, starting from a Hungarian data item it is possible to find its equivalent in Slovak by implicitly using the existing one-to-one relations between those two data items and their German counterpart.

Another important feature of the Ephras phraseology database application is a so-called description model. The description model for phrases in all four languages has been developed according to the latest phraseology principles and includes the following: basic form (i.e., the content of a phrase), meanings, style, grammar, collocation, pragmatics, examples, variants, keywords, synonyms, categories and multilingual comment.

From the user point of view the application has the following properties. The user can search within the database using any of four languages as the starting language. The search results are presented in a list form where the user can click on a particular search result and obtain the full description of the data item. The links to the related data items (e.g., equivalents) are included in the data item description.

The rest of the paper is organized as follows. The next section discusses the aspects of representing the phraseology data from the data modeling point of view. The subsequent section describes in details the application architecture

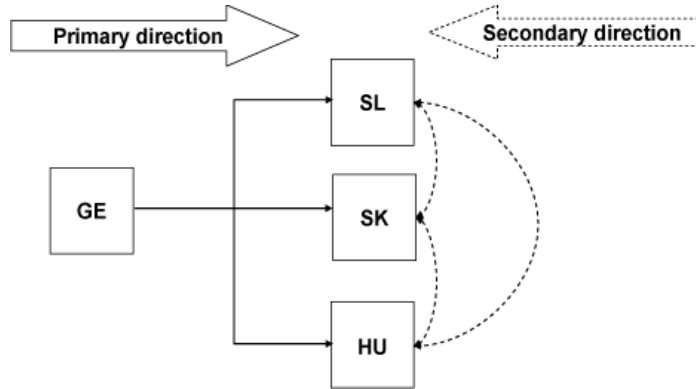


Fig. 1. Directions of Equivalence Relation

and the influence of the chosen data representation on that architecture. Finally, we give a number of conclusions and pointers for the future work.

2 Data Representation

To efficiently represent the phraseology data it was necessary to look closely on some of its features. Here is a list of some specific properties of the phraseology data:

- Data is structured only to a certain extent, i.e., there are data fields that have a different structure for different data items. For example, the meanings field contains typically textual content that possesses a varying internal structure - a single meaning or a list of meanings. Another example includes the multilingual comment field, where the content can be decorated, thus including text in bold or in italics.
- Whenever a data field has an internal structure as it is the case with the meanings or multilingual comment field, the ordering of elements within this internal structure is important and embodies semantic significance. For example, if a meanings field contains a list of meanings then the ordering of these meanings within the list possesses a certain denotation and needs to be maintained.
- Data items are interrelated by means of typed relations such as the equivalence relation discussed above. The equivalence relation is an ordered relation with varying arity and dimensions, e.g., the relation can be a one-to-one, an arbitrary one-to-many relation or even a many-to-many relation between meanings of data items. Additionally, data items can be involved in a co-called variance relation (e.g., a phrase is a variant of another phrase in the same language).

From the data modeling point of view, data with such properties is referred to as semi-structured data. Typically, semi-structured data is irregularly, partially or implicitly structured. Further, for such data there is no a-priori or rigid database schema but only a so-called a-posteriori data guide can be identified [1,2]. Obviously, the phraseology data in question can be classified as semi-structured data.

Generally, semi-structured data is modeled as a labeled graph [1]. The nodes represent data items, have unique identifiers and can be either atomic or composite. Composite data items are related with other data items by means of labeled edges, where labels represent the relation types. A simple model representing a couple of phraseology data items can be seen in Fig. 2.

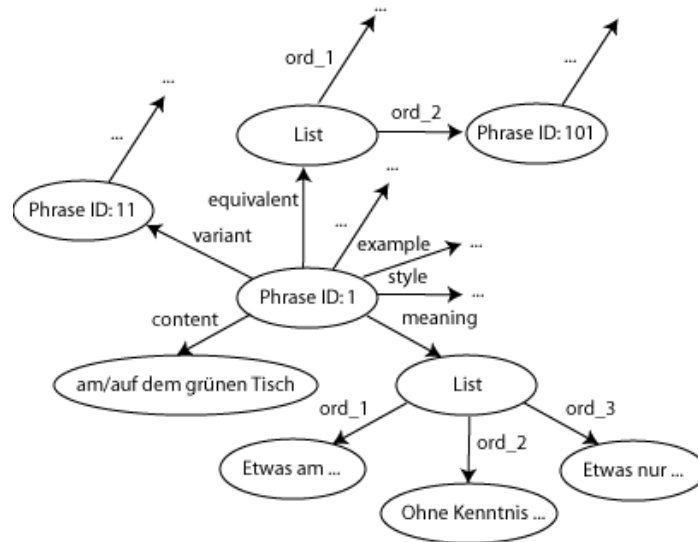


Fig. 2. Model of Semi-structured Phraseology Data

Recently, through the emergence of the Web and the related mark-up technologies, such as eXtensible Markup Language (XML), the latter evolved to a de-facto standard for managing semi-structured data [3,4]. An XML document is a hierarchy of elements with ordered sub-elements. Each element has a name (also referred to as label or tag). The basic XML model is a labeled ordered tree where labels represent node names. Edges are always directed (to preserve the tree order) and do not have labels. Additionally, XML supports a referencing mechanism between nodes, which basically facilitates modeling of arbitrary graphs. In this way semi-structured data might be represented by means of XML documents. An excerpt from an XML document encapsulating the above depicted phraseology data items is shown in listing 1.

```

<phrases>
  <phrase id="1">
    <content>am/auf dem grünen Tisch</content>
    <meanings>
      <meaning>Etwas am ... </meaning>
      <meaning>Ohne Kenntnis ... </meaning>
      <meaning>Etwas nur ... </meaning>
    </meanings>
    <style > ... </style>
    <examples>
      <example > ... </example>
    </examples>
    ...
  </phrase>
  <phrase id="11" variant="1">
    ...
  </phrase>
  <phrase id="101" equivalent="1">
    ...
  </phrase>
  ...
</phrases>

```

Listing 1. XML Document Encapsulating Phraseology Data Items

3 Implications of XML on Application Architecture

The architecture of the Ephras phraseology database application closely follows the well-known three-tiered architecture of user-oriented database applications (see Fig. 3). The three tiers are:

- User interface module that manages the user interaction and presents the data items to the user.
- Application logic module which implements the core application functionality by representing the data items and the operations that the user can perform on these data items (e.g., get an equivalent, get a variation of a phrase, etc.). This functionality is supported in a standard object-oriented manner, i.e., as a collection of interacting objects. Additionally, this module provides a bridge to the underlying data management module.
- Data management module which abstracts the access to an external XML-based database system by means of a programmatic interface. In addition, the external XML-based database system manages the XML representation of the phraseology data items.

Using XML for data management in the Ephras phraseology database application has a number of important aspects. First of all, the application deals

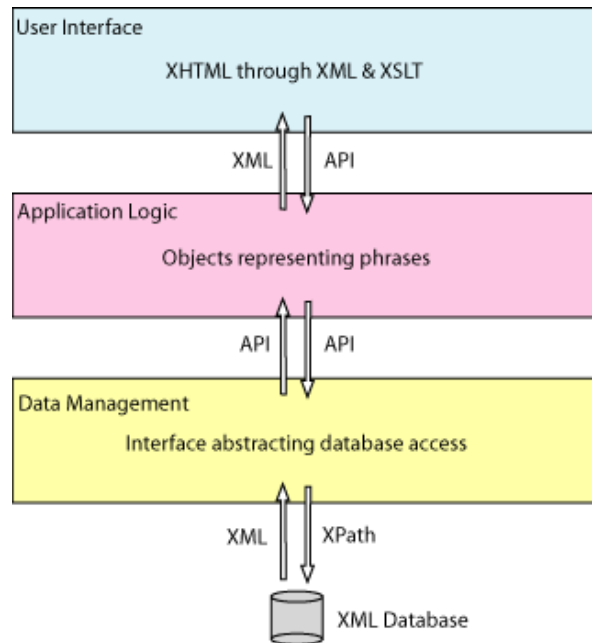


Fig. 3. Architecture of Phraseology Database

with the phraseology data from four different languages, namely German, Slovak, Slovenian and Hungarian. These four languages contain different characters, which are encoded using a particular character encoding schema. For example, German characters are encoded using ISO 8859-1 character set (Latin-1 or West European encoding). On the other hand, the characters from the remaining languages are encoded using ISO 8859-2 character set (Latin-2 or Central and East European encoding). Thus, the only possibility to combine characters from those four languages in a single XML document is to encode them using ISO 10646 Unicode character set. Technically, this does not constitute a problem, since XML documents might be encoded using Unicode character set. Additionally, all XML documents support UTF-8 and UTF-16 Unicode encodings, which define how to encode Unicode characters in a space-saving manner. In the Ephras phraseology database application we decided to use UTF-8 encoding for that purpose.

The second important aspect of using XML in the Ephras phraseology database application is related to communication between the data management module and the underlying XML-based database system. Obviously, the language for querying the database must be an XML query language. In this application the chosen language is XPath query language. XPath is a simple query language that works directly with the underlying tree-based model of an XML document supporting queries that retrieve subtrees of the whole XML tree. Thereby, different

matching criteria can be applied, such as element-based matching criteria (e.g., give me all phrase elements), attribute-based matching criteria (e.g., give me all phrase elements that have a certain attribute with a certain value) or content-based matching criteria (e.g., give me all phrase elements with a certain word in the content). For example, the first query in Listing 2 retrieves all phrases from the database and the second query retrieves all phrases that contain word 'Tisch'.

```
/phrases/phrase  
/phrases/phrase[content[contains(., 'Tisch')]]
```

Listing 2. XPath Queries for Retrieving Data Items

Finally, the third aspect of XML in the Ephras phraseology database application is related to the user interface module. Originally, XML is specified as a meta-document format that can be used to define families of document formats. Definition of presentation instructions for such document families is not a part of XML specification and is defined elsewhere - namely by a number of so-called style-sheet specifications. Basically, a style-sheet is a separate document which defines how a certain XML document should be presented. Currently, Cascading Style Sheets (CSS) and eXtensible Stylesheet Language - Transformations (XSLT) are typically used for that purpose. CSS is used to specify formatting instructions for XML documents whereas XSLT provides possibilities to transform an XML document to another XML document for which presentation instructions already exist. The best known example is transformation of arbitrary XML documents into HTML or XHTML documents, which can be subsequently presented using a standard Web browser. In this application we have chosen the latter approach and thus transform XML documents into XHTML documents and present them in a Web browser to the user.

4 Conclusion and Future Work

In this paper we presented the Ephras phraseology database application, which is a database application for management of phraseology data in four different European languages. For the purpose of implementing this application it was important to examine the defining features of such phraseology data. The most important technical result of this examination is the conclusion that phraseology data should be classified as semi-structured data. Since XML is a de-facto standard for management of semi-structured data today, applying XML database technology for implementing the application was an obvious choice. The subsequent discussion of a number of aspects of XML, such as querying facilities or presentation of the data provides an insight in a number of implementation issues.

Currently, the application is still in the development phase. The XML database, the data management module as well as the application logic module are already implemented. The user interface module is still under development. The first version of a complete system will be available in the beginning of 2006.

References

1. S. Abiteboul. Querying semi-structured data. In *Proceedings of the 6th International Conference on Database Theory - ICDT '97*, pages 1–18, 1997.
2. P. Buneman. Semistructured data. In *PODS '97: Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 117–121, New York, NY, USA, 1997. ACM Press.
3. R. Goldman, J. McHugh, and J. Widom. From semistructured data to xml: Migrating the lore data model and query language. In *Proceedings of the 2nd International Workshop on the Web and Databases (WebDB '99)*, 1999.
4. V. Vianu. A web odyssey: from codd to xml. In *PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 1–15, New York, NY, USA, 2001. ACM Press.