Interoperability of an 18th century Italian-Latin-

Petra Bago, Damir Boras
Department of information and communication sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{pbago, dboras}@ffzg.hr

Summary

The digitization process of historical texts is increasing in the last several decades, thereby developing the field of digital humanities. However, the digitization projects are usually isolated to the project teams, universities and institutes, but also to individuals developing the project. There was a lack of communication between the community members, which resulted in resources, tools and systems that are not able to exchange information.

Consequently, recently it was demonstrated that there is a demand for standardization of technologies, but also of all processes of development. Interoperability emerges as the key concept at this phase of digital humanities, which aims to facilitate communication of data.

To enable semantic interoperability of historical dictionaries, they have to be encoded using some standard. In this paper we present the encoding of della Bella's trilingual 18th century dictionary entries using a TEI (Text Encoding Initiative) encoding scheme for dictionaries. The dictionary is a trilingual dictionary containing Italian, Latin and Croatian language.

Keywords: historical dictionaries, interoperability, Text Encoding Initiative, encoding, digital humanities

Introduction

The digitization process of historical texts is increasing in the last several decades, thereby developing the field of digital humanities. However, the digitization projects are usually isolated to the project teams, universities and institutes, but also to individuals developing the projects. There was a lack of communication between the community members, which resulted in resources, tools, and systems that are not able to exchange information.

Consequently, recently it was demonstrated that there is a demand for standardization of technologies, but also of all processes of development. Interoperability emerges as the key concept at this phase of digital humanities, which aims to facilitate communication of data.

Interoperability defined

In order to overcome the lack of communication between language data, and achieve interoperability of language resources, two big international projects joined forces. One of the projects is a USA project called *Sustainable Interoperability for Language Technology* (SILT). Its 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 (SILT, 2014). The second project comes from Europe and is called *Fostering Language Resources Network* (FLaReNet). Its goal is to develop a common vision of the language resources and language technologies, and foster a European strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide (FLaReNet, 2014).

The aim of this international cooperation is to involve community members in the field of language technology, but also those working in related fields, to collaborate. The collaboration should create a consensus related to sharing data and technologies for language resources and applications, working towards the interoperability of existing data and, where possible, to promote standards for markup and resource creation (Ide et al., 2009).

One of the main achievements of this international cooperation is a definition of the term "interoperability" within the field of language technologies (Ide et al., 2009).

According to Ide et al. (2009) interoperability can broadly 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 generally divided in two types: syntactic interoperability and semantic interoperability. Syntactic interoperability aims at enabling communication and data exchange, relying on specific data formats, communication protocols and the like. It is important that information is exchanged, but there is no guarantee that interpretation of this information will be the same. On the other hand, with semantic interoperability, systems have the ability to automatically interpret exchanged information meaningfully and accurately in order to produce useful results via compliance to a common information exchange reference model. It is important that interpretation of exchanged information is the same on both sides of the communication (Ide et al., 2009).

To enable interoperability of language resources, the focus 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. Hence, syntactic interoperability in the context of language resources can be defined as the ability of different systems to process exchanged data either directly or via trivial conversion. On the other hand, semantic interoperability in the context of language resources can be defined as the ability of systems to interpret exchanged information in meaningful and consistent way. In that sense, language resources

have to focus more on semantic and not on syntactic interoperability (Ide et al., 2009).

Semantic interoperability is the reason why we decided to use TEI (Text Encoding Initiative) encoding scheme for dictionaries ("Dictionaries", 2013). The TEI Consortium develops and maintains the Guidelines (TEI P5, 2014), intended for everyone who is generating and/or processing textual resources in digital form. The Guidelines are recommendations of the Consortium on how to encode implicit features of textual resources, thereby making these features explicit. Once the structure or some other feature of a text is explicitly encoded, later processing by computers is easier. Text encoding with a predefined scheme enables data exchange without or with minimal information loss, and correct interpretation of the information. Since the Guidelines are based on XML, they are independent of computer programs and operation systems. The Guidelines can be applied to all natural language texts, without any limitations to textual form or content, as well as digitized and digitally born texts ("About these Guidelines", 2014).

The 9th chapter of the Guidelines defines tags that can be used to encode dictionary entries, as well as other lexical resources, such as glossaries ("Dictionaries", 2014). We consulted this chapter when encoding entries from an old Croatian printed dictionary with a very complex entry structure. Our research is conducted on the first volume of a second edition of a trilingual dictionary "Dizionario italiano-latino-illirico" (Italian-Latin-Croatian dictionary) compiled by Ardelio della Bella and printed in Dubrovnik in 1785 (della Bella, 1785). This dictionary, among several other important Croatian historical dictionaries, was digitized as part of the scientific projects "Croatian dictionary heritage and dictionary knowledge representation"[1] and "Croatian dictionary heritage and Croatian European identity"[2] funded by the Croatian Ministry of Science and Technology. The digitization process of the dictionaries is not the scope of this paper.

The process of encoding digitized dictionary entries is performed manually, which is time consuming and tedious. Since the process is a manual task, it paves a way for an encoder to introduce inconsistencies. If the process of encoding could be (semi)automated, it would reduce cognitive load in encoders and time spent on the task. However, to (semi)automate the encoding process, first we have to analyze the structure of the entries of a chosen dictionary, and encode the content, preferably using a standard. The scope of this paper is to

---

[1] Boras, Damir (principal researcher). Hrvatska rječnička baština i prikaz rječničkoga znanja; Znanstveni projekt, Zagreb, 2003-2006. http://zprojekti.mzos.hr/public/c2prikaz_det.asp?cid=1&psid=25&ID=451 (25.8. 2014.)

[2] Boras, Damir (principal researcher). Hrvatska rječnička baština i hrvatski europski identitet; Znanstveni projekt, Zagreb, 2007-2013. http://zprojekti.mzos.hr/public/c2prikaz_det.asp?cid=1&psid=25&ID=451 (25.8. 2014.).

describe this first step i.e. to present the results of the manual encoding of della Bella's dictionary entries.

About della Bella's dictionary

The dictionary was intended for Italian Jesuit missionaries to help them spread the faith in a national language i.e. Croatian language, but also other Slavic languages. For this reason a Croatian grammar can be found inside the dictionary preamble. The dictionary contains 899 pages and consists of two parts. The first part is a preamble written in Italian language on 54 pages. The second part is the dictionary, containing around 19,000 headwords. The dictionary is printed in two volumes: the first volume contains the preamble and the dictionary part from letters A to H, while the second volume contains the dictionary part from letters I to Z. For the first time in Croatian lexicography, della Bella's dictionary contains examples of uses of headwords in various literary works and oral literature (Bago & Boras, 2012).

Encoding dictionaries

Two main problems arise while encoding dictionaries. One problem is related to the structure of the entry, while the other problem is related to the information found within the entry. The structure of entries varies widely among and within dictionaries. Since the encoding scheme is supposed to be suitable for various entry structures, it allows every element of the scheme to appear anywhere in a dictionary entry. However, there are dictionaries that have consistent structure, thus the need for the scheme to support such cases. Since della Bella's dictionary has a complex, but still consistent structure, we decided to encode the entries with the element <entry>, which is intended for consistent and regular dictionary entries[3].

The second problem arising while encoding dictionaries is the problem of representation of information within an entry. Most information in dictionaries is implicit or compressed. Therefore, the encoder must decide whether to capture the precise typographic form of the source text or the underlying structure of the information it presents. Some encoders may find it important to be faithful to an original printed version. On the other hand, others may find only lexical information important, thereby changing, adding or deleting the information contained in the printed form. Both approaches are also possible to combine[4].

We decided to encode old Croatian dictionaries by keeping the information contained in the printed form as content of the elements, while additional information that cannot be explicitly found in the original is encoded as attributes

---

[3] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-entry.html

[4] TEI Consortium (ur.). "9. Dictionaries".TEI P5: Guidelines for Electronic Text Encoding and Interchange. (Verzija 2.6.0). TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (10. 7.2014.).

of the elements. That way we separate the original content from the new, added content.

Scheme of della Bella's dictionary entries

In this section we present the scheme structure of entries found in della Bella's dictionary.

The root element for an entry is the element <entry>. The element <form> groups all the information on the written and spoken forms of one headword[5]. The attribute @type classifies what form is encoded, i.e. simple, compound, etc. The value for the headword itself is "lemma", which denotes that the headword is in its elementary form. Additional attribute is used to define the language of the headword as important information for bilingual and multilingual dictionaries. Since della Bella's dictionary has only an orthographic form of the headwords, the element <orth> is used[6].

All punctuation characters are explicitly encoded with the element <pc> (punctuation character)[7]. There is a space before every punctuation character that is not part of an abbreviation. There is no space before a punctuation character that is part of an abbreviation.

According to the description given above, the headword "Abate" is encoded in the following way:

```
<form type="lemma" xml:lang="it">
<orth>Abate</orth>
<pc>.</pc>
</form>
```

Within a dictionary entry, a headword can additionally be found in a form other than the usual dictionary form. Generally, not whole words can be found but its suffixes. The reason for this is saving space. These suffixes are encoded with the value "inflected" of the @type attribute. A suffix "ta", which represents a suffix of the singular noun in genitive of the translation "Opat", is encoded as following:

```
<form type="inflected">
<gramGrp>
<case value="genitive"/>
<number value="singular"/>
 ta
<pc>.</pc>
</gramGrp>
</form>
```

---

[5] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-form.html

[6] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-orth.html

[7] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-pc.html

Three new elements can be found above. The element <gramGrp> (grammatical information group) groups morpho-syntactic information about a lexical item[8]. The element <case> contains grammatical case information for a given form, and its value of the case in an attribute @value[9]. In this case the value of the attribute @value is "genitive". The element <number> indicated grammatical number associated with a form and its value of the attribute @value[10]. In this case the value is "singular". The elements <case> and <number> are not explicitly given, but are implicitly given via the suffix. Therefore these elements are added as empty elements. As we mentioned before, the content of the elements are only those parts found in the original printed version.

Furthermore, with the element <form> a compound can be encoded that is formed from simple lexical items, one being the headword. It is encoded with the value "compound" of the @type attribute. The element <oRef/> (orthographic-form reference) is used to indicate a reference to the orthographic form(s) of the headword[11]. A compound "Dignita d'Abate" of the headword "Abate" is encoded as following:

    <form type="compound">
    <orth>Dignità d'<oRef/></orth>
    <pc>.</pc>
    </form>

In the printed dictionary following a suffix for a singular noun in genitive is an abbreviation for the morphological gender of a lexical item, which we encode with the element <gen> and the attribute @value[12]. Apart from an abbreviation for the gender, we added an element for the part of speech (<pos>) and the value of the assigned part of speech is recorded with the attribute @norm[13]. The element <pos> is an empty element, since it is not explicitly given in the printed dictionary. An abbreviation for a masculine noun is encoded as following:

    <gramGrp>
    <pos norm="noun"/>
    <gen value="m">m<pc>.</pc></gen>
    </gramGrp>

The element <sense> groups together all information relating to one word sense in a dictionary entry, i.e. definitions, examples and translations. With the attrib-

---

[8] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-gramGrp.html

[9] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-case.html

[10] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-number.html

[11] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-oRef.html

[12] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-gen.html

[13] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-pos.html

ute @n one can encode the number of different senses found within an entry[14]. Inside this element, the element <cit> (cited quotation) with a value "translation" of attribute @type contains a translation of the headword, while the value "example" of the same attribute contains an example of the headword[15]. A mandatory element within the <cit> is <quote>, which contains a phrase or passage attributed by author of the dictionary to some agency external to the text[16]. A translation of the entry "Abate" on Latin is encoded as following:

    <cit type="translation" xml:lang="la">
    <quote>Abbas
    <pc>,</pc></quote>
    </cit>

An example of Croatian translations are encoded as following:

    <cit type="translation" xml:lang="hr">

while in examples there is a source of the example given, tagged with the element <bibl>[17]. Since there is no example in the entry "Abate", here we present an example found in the entry "Anno":

    <cit type="example" xml:lang="hr">
    <quote>Evo gre pet godin', dàsam gne sluga ja <pc>;</pc></quote>
    <bibl>Sciſc<pc>.</pc></bibl>
    </cit>

Sometimes dictionaries contain unclassifiable piece of information to guide sense choice, which is encoded with the element <usg> and attribute @type with value "hint"[18]. In the following example the lexicographer describes the headword "Affrica" more precisely by stating that it is "one of the four parts of the World":

    <usg type="hint">
    una delle quattro parti del<lb/>Mondo <pc>.</pc>
    </usg>

Dictionaries have many references to other entries within the same dictionary, which are encoded with the element <xr> (cross-reference phrase) and attribute @type with value "see" 19. Within this element, there is the element <ref> (reference) that defines a reference to another location, in this case another headword20. To define the location of the referenced entry, we use the attribute

---

[14] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-sense.html

[15] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-cit.html

[16] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-quote.html

[17] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-bibl.html

[18] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-usg.html

[19] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-xr.html

[20] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-ref.html

@target with the value of the location of the entry, as in the following example of the entry "Abbarbicare":

```
<xr type="see">V<pc>.</pc>
<ref target="#barbare">Barba-re</ref>
<pc>.</pc>
<ref target="#radicare">Radicare</ref>
<pc>.</pc>
</xr>
```

In this case, the target entries have to have an @xml:id with a unique value within the XML document in the <form> element. An example of the entry "Barbare" referenced in the entry "Abbarbicare" is encoded as following:

```
<form type="lemma" xml:lang="it" xml:id="barbare">
<orth>Abate</orth>
<pc>.</pc>
</form>
```

Finally, we encode line breaks (<lb/>), column breaks (<cb/>) and page breaks (<pb/>).

Below we present the whole content of the entry "Abate" as found in the della Bella's dictionary:

Abate . *Abbas , tis . m.* Opat , ta . m. Igu-|men , ena . m. Dignità d'Abate . Opat-|ſtvo , va . n. Igumenſtvo , tva . m.

and the encoding of the corresponding digitized dictionary entry:

```
<entry>
<form type="lemma" xml:lang="it">
<orth>Abate</orth>
<pc>.</pc>
</form>
<sense n="1">
<cit type="translation" xml:lang="la">
<quote>Abbas
<pc>,</pc></quote>
<form type="inflected">
<gramGrp>
<case value="genitive"/>
<number value="singular"/>
   tis
<pc>.</pc>
</gramGrp>
</form>
<gramGrp>
<pos norm="noun"/>
<gen value="m">m<pc>.</pc></gen>
</gramGrp>
```

```
</cit>
<cit type="translation" xml:lang="hr">
<quote>Opat
<pc>,</pc></quote>
<form type="inflected">
<gramGrp>
<case value="genitive"/>
<number value="singular"/>
   ta
<pc>.</pc>
</gramGrp>
</form>
<gramGrp>
<pos norm="noun"/>
<gen value="m">m<pc>.</pc></gen>
</gramGrp>
</cit>
<cit type="translation" xml:lang="hr">
<quote>Igu-<lb/>men
<pc>,</pc></quote>
<form type="inflected">
<gramGrp>
<case value="genitive"/>
<number value="singular"/>
  ena
<pc>.</pc>
</gramGrp>
</form>
<gramGrp>
<pos norm="noun"/>
<gen value="m">m<pc>.</pc></gen>
</gramGrp>
</cit>
</sense>
<sense n="2">
<form type="compound">
<orth>Dignità d'<oRef/></orth>
<pc>.</pc>
</form>
<cit type="translation" xml:lang="hr">
<quote>Opat-<lb/>ſtvo
<pc>,</pc></quote>
<form type="inflected">
```

```
<gramGrp>
<case value="genitive"/>
<number value="singular"/>
    va
<pc>.</pc>
</gramGrp>
</form>
<gramGrp>
<pos norm="noun"/>
<gen value="n">n<pc>.</pc></gen>
</gramGrp>
</cit>
<cit type="translation" xml:lang="hr">
<quote>Igumenſtvo
<pc>,</pc></quote>
<form type="inflected">
<gramGrp>
<case value="genitive"/>
<number value="singular"/>
    tva
<pc>.</pc>
</gramGrp>
</form>
<gramGrp>
<pos norm="noun"/>
<gen value="m">m<pc>.</pc></gen>
</gramGrp>
</cit>
</sense>
</entry>
```

Conclusion

To enable semantic interoperability of digitized historical dictionaries, the dictionaries have to be encoded using some standard. In this paper we present the successful encoding of della Bella's 18th century dictionary entry using a TEI (Text Encoding Initiative) encoding scheme for dictionaries. We consider the TEI scheme an adequate method for encoding dictionaries, since it is a *de facto* standard for encoding textual resources in digital form, spread within the field of digital humanities. Additionally, we consider it adequate, considering it enables us to encode all the necessary information found in dictionaries, regardless of its complicated structure.

The process of encoding digitized dictionary entries is a manual task, which is time consuming and tedious, and paves a way for an encoder to introduce in-

consistencies. If the process of encoding could be (semi)automated, it would reduce cognitive load in encoders and time spent on the task. We could further apply natural language processing methods (such as machine learning) to try to speed up the process of encoding the structure of the dictionary entries.

By using the TEI scheme, we could additionally link the entries and concepts to external resources, such as online encyclopaedias (i.e. the Croatian Encyclopaedia of the Miroslav Krleža Institute of Lexicography[21] or the free-content Wikipedia[22]), online dictionaries (i.e. Croatian language portal[23]), Linked Data[24] or DBpedia[25].

References

"About These Guidelines" in TEI P5: Guidelines for Electronic Text Encoding and Interchange. TEI Consortium (ed.).. http://www.tei-c.org/Guidelines/P5/ (10. 7.2014.).

Bago, Petra; Damir, Boras. A framework for consolidating most important digitized Croatian dictionaries from 1595 to 1945 // 17. International Information Technology and Journalism conference. Dubrovnik, 2012.

"bibl" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-bibl.html (25.09.2014.).

Boras, Damir (principal researcher). Hrvatska rječnička baština i prikaz rječničkoga znanja; Znanstveni projekt, Zagreb, 2003-2006. http://zprojekti.mzos.hr/public/c2prikaz_det.asp?cid=1&psid=25&ID=451 (25.8. 2014.).

Boras, Damir (principal researcher). Hrvatska rječnička baština i hrvatski europski identitet; Znanstveni projekt, Zagreb, 2007-2013. http://zprojekti.mzos.hr/public/c2prikaz_det.asp?cid=1&psid=25&ID=451 (25.8. 2014.).

"case" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-case.html (16.09.2014.).

"cit" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-cit.html (21.09.2014.).

The Croatian Encyclopaedia of the Miroslav Krleža Institute of Lexicography. http://www.enciklopedija.hr/ (05.09.2014.).

The Croatian Language Portal. http://hjp.novi-liber.hr/ (05.09.2014.).

DBpedia. http://wiki.dbpedia.org/ (05.09.2014.).

della Bella, Ardelio. Dizionario italiano-latino-ilirico. Nella Stamparia Privilegiata, Prima edizione Ragusea, 1785.

"Dictionaries" in TEI P5: Guidelines for Electronic Text Encoding and Interchange ed. The TEI Consortium (ed.). 2014. V. 2.6.0. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html.

"entry" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-entry.html (15.09. 2014.).

eXtensible Markup Language (XML). http://www.w3.org/XML/ (10.7.2014.).

"form" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-form.html (15.09. 2014.).

Fostering Language Resources Network (FLaReNet). https://www.flarenet.eu/ (5. 9. 2014.).

"gen" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-gen.html (18.09.2014.).

---

[21] http://www.enciklopedija.hr/

[22] https://en.wikipedia.org/wiki/Main_Page

[23] http://hjp.novi-liber.hr/

[24] http://linkeddata.org/

[25] http://wiki.dbpedia.org/

"gramGrp" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-gramGrp.html (16. 09.2014.).

Ide, Nancy; Pustejovsky, James; Calzolari, Nicoletta; Soria, Claudia. The SILT and FLaReNet International Collaboration for Interoperability. Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009, str. 178-181.

Linked Data – Connect Distributed Data across the Web. http://linkeddata.org/ (05.09.2014.).

"number" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-number.html (18.09. 2014.).

"oRef" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-oRef.html (18.09. 2014.).

"orth" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-orth.html (15.09.2014.).

"pc" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-pc.html (15.09.2014.).

"pos" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-pos.html (18.09.2014.).

Pustejovsky, James; Ide, Nancy. "What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability," 2010. http://www.cs.vassar.edu/~ide/papers/ ICGL10.pdf (5. 9. 2014.).

"quote" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-quote.html (25.09. 2014.).

"ref" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-ref.html (28.09.2014.).

"sense" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-sense.html (18.09. 2014.).

Sustainable Interoperability for Language Technology. http://www.anc.org/LAPPS/SILT/index. html (5.9.2014.).

TEI Consortium (ed.). TEI P5: Guidelines for Electronic Text Encoding and Interchange. (Version 2.6.0). TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (10. 7.2014.).

"usg" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-usg.html (28.09.2014.).

Wikipedia, the Free Encyclopedia. https://en.wikipedia.org/wiki/Main_Page (05.09.2014.).

"xr" in TEI P5. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-xr.html (28.09.2014.).