An XML Version of Turkish Dictionary

Emrah Özcan

An XML Version of Turkish Dictionary[1,2]

Emrah Özcan
Yıldız Technical University, Istanbul, Turkey - eozcan@yildiz.edu.tr

In order to anchor international or multinational lexicographic projects on existing Turkish dictionaries, we should have a common understanding of the way we make reference resources available, as is the case for the digital version of our Turkish dictionary project. Although our work has been done on digitizing Turkish dictionaries, both old Turkish dictionaries and current versions of it, these few examples do not follow a standard way of encoding the source file. In order to overcome this obstacle, during the Lexical Data Masterclass in Berlin, on December 4-8, 2017, I worked on an XSLT transformation document to process an existing dictionary into an output conformant to the TEI standard. Seeing that almost all current Turkish dictionaries give the same category of lexical information in a very similar page layout, this XSLT could work on other digitalized or OCRized Turkish dictionaries. Even if they do not have a digital version, GROBID based projects can easily transform OCRized PDFs into digital file format like the works of GROBID-Dictionaries (Khemakhem et al. 2017).

The entry structure that I have been using is taken from the Turkish Dictionary published by the *Turkish Language Institute* (abbreviated as "TDK" in Turkish) which acts as the official authority on the Turkish language (without any enforcement power) and it contributes to linguistic research on Turkish. Almost all mainstream Turkish dictionaries, such as the ones published by Dil Derneği – Language Foundation (Dil Derneği Türkçe Sözlük, 2005) and Arkadaş Publishing (Püsküllüoğlu, 2004), share similarities by means of the lexical information given in the dictionary and also the page layout. The only difference is the order of information given in the microstructure. For example, the order of the etymological and phonological information is different; TDK (Turkish Language Institution) dictionary gives phonological information before the etymological information and the other dictionaries do the opposite; they first give the etymological information and then the phonological information. The rest of the page layout gives the same lexical information in the same order. Therefore, the same XSLT file could handle all Turkish dictionaries to do such a transformation for an appropriate TEI standard. I should also thank Laurent Romary for his invaluable help while I was creating the XSLT file during the Lexical Data Masterclass.

---

[1] This work was presented at the Lexical Data Masterclass Participants' Symposium co-organized by DARIAH, the Berlin Brandenburg Academy of Sciences (BBAW), Inria and the Belgrade Center for Digital Humanities, with the support of the German Ministry of Education and Research (BMBF), Clarin and DARIAH-DE, which took place in Berlin from 4 to 8 December 2017.

[2] An initial version of this work is first published in the Lexical Data Masterclass website: https://digilex.hypotheses.org/275

In the XSLT file, we first used the **<teiheader>** to declare the necessary information about the dictionary. This section contains information about the title of the dictionary, the author, and the publication statement which also contains copyright information. TEI-header also includes the bibliographic information about the dictionary that includes the edition, publisher and publishing date.
After using the TEI header, we start the transformation process. We begin with the first parent element under the root element and its children element for each entry. For this, we use **<xsl:template>** with "**@match**" attribute to put all the entries under the **<entry>** tag. Inside this xsl:template, we use **<xsl:template>** again but this time we use the "**select**" attribute to select all the words in the dictionary.

```
<xsl:template match="madde">
<entry>
<xsl:apply-templates select="sozcuk"/>
</entry>
</xsl:template>
```

Now, we have all our entries inside the **<entry>** element. In each entry, we have lemmas, and lemmas are tagged with **<form>** element with "**type**" attribute named as "**lemma**".

```
<xsl:template match="sozcuk">
<form type="lemma">
<orth>
<xsl:apply-templates/>
</orth>
</form>
</xsl:template>
```

Since Turkish is an agglutinative language, certain suffixes change the ending of the word and Turkish dictionaries provides this information. **<form type="inflected">** is used to reflect this change in the XML file. **<orth>** is used after the the **<form>** tag since this change is related to the orthography of the word.

```
<xsl:template match="degisim">
<form type="inflected">
<orth>
<xsl:apply-templates/>
</orth>
</form>
</xsl:template>
```

For grammatical information like part of speech, **<gramGrp>** and **<pos>** is used. For additional grammatical information **<subc>** is used, like cases.

```
<xsl:template match="dilbgs">
<gramGrp>
<pos>
<xsl:apply-templates/>
</pos>
</gramGrp>
</xsl:template>

<xsl:template match="durum">
<subc>
<xsl:apply-templates/>
</subc>
</xsl:template>
```

Next is the etymological information. For etymological information **<etym>** is used. **<etym>** tags can also be used with **<lang>** tags where you can specify the language origin.

```
<xsl:template match="kok">
<etym>
<lang>
<xsl:apply-templates/>
</lang>
</etym>
</xsl:template>
```

For the examples used in the dictionary, **<cit type="example">** is used. All examples used in TDK dictionary is taken from literary pieces, like novels, poems etc. **<quote>** and **<bibl>** tags are being used after **<cit>** since they refer to a bibliographic reference.

```
<xsl:template match="ornek">
<cit type="example">
<quote>
<bibl>
<xsl:apply-templates/>
</bibl>
</quote>
</cit>
</xsl:template>
```

The author information for the example is also given in the dictionary by using the **<author>** tag.

```
<xsl:template match="ornek_kaynak">
<author>
<xsl:apply-templates/>
</author>
</xsl:template>
```

At the end of the entry, TDK gives related entries that can be related to the current entry. In order to show this information, we used **<re>** and the **“type”** attribute for specifying the degree of relatedness, **“collocation”** or **“proverb”** etc.

```
<xsl:template match="atasozu_deyim_birlesikfiil">
<re type=“proverb-idiom-compoundVerb“>
<xsl:apply-templates/>
</re>
</xsl:template>
<xsl:template match="birlesiksoz">
<re type="collocation">
<xsl:apply-templates/>
</re>
</xsl:template>
```

References

(1) Dil Derneği Türkçe Sözlük. (2005). Dil Derneği, Ankara.

(2) Khemakhem, M., Foppiano, L., & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. in electronic lexicography, *eLex 2017*, Sep 2017, Leiden, Netherlands. 〈hal-01508868v2〉

(3) Püsküllüoğlu, A. (2004). *Arkadaş Türkçe Sözlük*. Doğan Kitap, İstanbul.

(4) *Türkçe Sözlük*. (1998) (9th edition). Türk Dil Kurumu Yayınları, Ankara.