Abstract
In this paper, we identify the need for a standardized formalism for the structured XML dictionaries of endangered Uralic languages in the Giella infrastructure. For this purpose, we have decided to use TEI formalism as it is a standardized way of representing data and its commonly used in the field of lexicography. This paper focuses on describing the issues and challenges faced in the conversion of the Giella XML into TEI. A full conversion scheme is introduced in this paper contrasting the peculiarities of the two XML formalisms. We incorporate the new TEI-based XML structure into our existing online dictionary system as an output format.

Key Words: endangered languages, XML-MediaWiki, TEI, Uralic languages

Introduction
This paper addresses dictionary-resource development for endangered, under-resourced languages in collaboration with an open-source infrastructure with a rule-based orientation as described in Moshagen et al. (2014). It then outlines advances in XML—MediaWiki synchronization of multilingual dictionaries (Hämäläinen & Rueter, 2018) and enhanced features for etymological and cognate resource work (Hämäläinen & Rueter, 2019) and automatic combination of concepts in multilingual dictionaries (Hämäläinen, Tarvainen & Rueter, 2018). Work with XML introduces a need for a standardized TEI (Text Encoding Initiative) formalism.

As noted in Czaykowska-Higgins (2014), XML structuring greatly benefits from international TEI standards developed since 1990s. Numerous applications bolster personal and professional usage of emerging technologies. Simultaneously, work addresses individual nodes and issues, e.g. etymology (Bowers & Romary, 2016), digitization (Maxwell & Bills, 2017), and endangered language resource development (Czaykowska-Higgins, ibid.). The utilization of TEI standard affords shared usage of tools and databases on many platforms as well as multiple possibilities for transformation, rendering and publication.

In the most recent update for TEI (29th January 2019), dictionary guidelines are characterized as catering towards human-oriented presentation. Although readily applicable to majority-language dictionary development, this practice may require tweaking for endangered and low-resourced languages.
Thus, this paper investigates alterations to the orientation in favor of rule-based language-technological infrastructures catering to low-resourced, endangered languages. This entails a strategy of TEI-compatibility, computer-legibility and facilitation of rule-based technologies.

1. TEI-compatibility observed in convertible shorthand XML tags, e.g. l = lemma.
2. Computer-legibility, delimited XML structure depth, unique element-type naming policy.
3. Rule-based description for minimal repetition and expenditures in language-resource development.

Our strategy is to join lexicographical and language-technology efforts for language (re)vitalization, e.g. click-in-text multilingual dictionaries, spellcheckers, etc. This means the introduction of stem-type and inflectional data in lemma-adjacent nodes, something outside the scope of TEI. The solution involves XSLT to formats addressed in Bánski et al. (2017).

**Method**

A great number of dictionaries in the Giella infrastructure (Moshagen et al., 2014) follow an XML structure that serves a purpose in the infrastructure itself. However, for external use such a format can be seen as troublesome due to insufficient documentation and standardization practices. The fact that these XMLs can be edited in a synchronized way in our MediaWiki environment (Rueter & Hämäläinen, 2017) makes it possible to include new XML formalisms without interfering with the existing Giella infrastructure. As our MediaWiki based online dictionary has been designed with the notion of multiple realizability of the data in different formats, adding a TEI support is just a matter of defining the correspondences between the Giella XML and the TEI standard.

The Giella XML dictionary structure is focused to address issues of machine-readability, minimal weight and reusability. To ensure machine-readability the depth of a given entry element does not exceed four, and element names can only be shared by same-depth elements.

The issue of minimal weight is addressed by establishing mnemonic one-, two- or three-letter element names, which are readily convertible to the TEI standard, but, which for purposes of light infrastructure are used as is in every-day code:

The e element stands for entry, this is the base of an entire word article. This element has both attribute and element content. The attribute information address matters of identity (in id) and exclusion from specific usage, i.e. exclude generally has the value fst (finite-state transducer), which means this particular article is not used in finite-state transducer generation. The minimal contents are one singular lg element (the lemma group element) and one or more mg elements (meaning group, i.e. sense group). The lg element can be preceded optionally by a map element, which contains attributes and values pertaining to original dictionary sources, and a rev-sort_key element, whose text content consists of the lemma or head word in reverse (right-to-left). The obligatory lg element may be immediately followed by a sources element with child elements referring to both source literature and parallel attestation of the lemma in other sources. The resources element data should, in fact, be directly associated with semantic meaning, and therefore in the future it will be moved to the appropriate lg and mg subelements.

The lg (lemma group) element has no attributes, but it does contain numerous child elements that can be directly associated with word form and not semantics. The two most prevalent child elements of the lg are the singular l (lemma) and stg (stem group) elements. These two elements provide information necessary
for the machine description. Other elements are optional but provide additional information useful in word
and word form recognition (audio, etymology, compg, mini_paradigm).

The head word text content of the l element is augmented by the presence of attributes. These consist, for
example, of pos (indicating part-of-speech), hid (homograph/homonym with values: Hom1, Hom2, …),
whereas the lemma or presentational form of one entry may be identical to that of another, but other
morphological forms or origin may distinguish them; only words from the same part-of-speech have
distinguishing hid attributes), type (e.g. common vs. proper noun, where common nouns are default and
proper nouns are shown with attributes), val (valency of verbs, with initial transitive vs. intransitive
marking). All of this information is used in the construction of the finite-state description of a given word
in the source language.

The stg element has no attributes, and the only child elements it may have are st (stem) elements. Each
individual st element has linguistically relevant text content representing a working morphological stem that
all word forms of the paradigm can be derived from in the Giella infrastructure. The attributes, in turn,
provide information on inflection type (both for the end user: machine and human), as well as additional
data revealing orthographic and language norm status. Since the Giella XML structure allows for
pluricentric documentation of a language, i.e. audio and orthographic representations for divergent places
in time and space, there are also varid (variant identifier) attributes with which to align audio, stem and even
possibly mini_paradigm content. The varid attribute is used in the st (stem) element whenever there are
more than one st element in the stg (stem group) element. This serves as a parallel backup to the principle
of “prefer first sibling when there are more than one to choose from”, which is necessary when the system
is expected to generate a single preferred word form.

The audio (this represents audio link information) element is optional. It may have a varid (variant identifier)
attribute to align it with an st sub-element in stg. Otherwise it has child elements with information on the
audio identifier, the reader, etc. Some of this information could be moved to a different location to minimize
the XML content.

In addition to dividing entries according to inflection, they are further divided by an etymological criterion.
Thus the etymology element only occurs once per entry, and it takes no attributes. It can have multiple
eyton and cognate child elements. The etymon element is of mixed content with attributes designating pos
(part-of-speech), algu_lekseemi_id (link information to the external etymological database Álgu), xml:lang
(639 ISO Language Code reference), in addition to lemmaID (lemma identifier) and stemID (stem
identifier). The cognate element has been used for crosslinking to other dictionaries in the multilingual

The compg (compound group) element is used for documenting compound words, derivations and
inflections alike. While the parent element has attributes drv (derivation which can spell out the
concatenation with the resulting part-of-speech) and type (values are: Cmp compound, Der derivation, Infl
inflection), there are ordered comp (compound) sub-elements containing link information. The comp
element has obligatory ord (order) attributes to establish constituent order (values: E1, E2,...) although in
most instances there are only two comp elements. The text content of the individual comp element is the
lemma or head word, which is then complemented by morpho-syntactic tags in an msd attribute, e.g. the
value N.Sg.Gen might tell us that the specific constituent is a noun appearing in the genitive singular. The
pos (part-of-speech) attribute here is simply a fallback for when there is no morphological analysis available,
but it is also used to implement crosslinking to the source lemma elsewhere in the dictionary. If the comp
The mini_paradigm element provides editors with an opportunity to give feedback on the paradigm produced by the finite-state transducers. These are legacy elements whose output will be used as tickets for prompting improvement in paradigm generation work. In generated pages transducer-produced mini-paradigms will, by default, show the content of the edited mini-paradigm, which can subsequently be turned off by adding an attribute exclude with the value aku (for the akusanat dictionaries). The mini_paradigm element has child elements in analysis and grandchild elements in wordform. The analysis element has an msd attribute providing morpho-syntactic analysis with tags separated by full stops. Since there are possibilities of multiple word forms, there may be more than one wordform element in which case a varid (variant identifier) attribute is necessary.

**Results**

The corresponding element to the e element in TEI is entry. As there is no direct correspondence for lg in TEI, the information stored in this elements is split into different parts of the TEI structure. The l element containing the lemma and part-of-speech is separated into two different tags: orth containing the lemma and pos under gramGrp containing the part-of-speech. The TEI gramGrp element also contains the inflectional information from the Giella stg and st elements under iType and cit.

The audio tags are moved to cit elements under form element. Mini_paradigm is expressed as a form element of infl type. The compg element expressing the compounds that constitute the lemma are split into cit elements that project a new dictionary entry structure to express the same information.

The mg level is moved to sense tags and the t elements containing the translations are nested as cit elements directly to under the entry tag. Finally, the xg tags containing the examples are expressed as cit elements containing quote elements in the TEI structure.
Figure 1: A Giella XML and TEI version of the same entry

Figure 1 shows the structural difference between the existing Giella XML and the TEI XML elaborated in this paper. Both formalisms are capable of representing the same data, but there is a difference in terms of compactness of the two.

**Discussion**

The Giella XML, despite its problems, caters for the need of machine readability and parsability. For this reason the XMLs can be widely used by different tools and services in the infrastructure. The TEI XML introduces more unnecessary complexity for machine readability as the foundations of its design seem to be in preserving the structure of a printed dictionary in a digital format.

However, as for the longevity and reusability of the dictionaries in the future, TEI provides better prospects due to the fact of documentation and standardization. This makes it possible to process TEI XMLs with a multitude of third party applications that provide support for the standard out of the box. Therefore, for us the gain of implementing the TEI formalism is in making the dictionary data available for export in a well-supported format for others to use.

There are a few discrepancies in the XML structure utilized in the Giella infrastructure and those set forth in the TEI standard. While the Giella XML structure caters to dictionary word form generation for multiple
reusability, the TEI standard appeals to visual presentation in paper-print and HTML dictionary pages. In practice, the Giella XML has been engineered to serve as a language-independent yet multilingual database, where source- and target-language data are stored in parallel structures, which would allow for language pair flip analysis and sanity checks. The TEI standard offers each individual dictionary project XML structuring possibilities that help guarantee presentation retention for any number of dictionary writing traditions, i.e. the convergence between the Giella XML structure and that of the TEI standard might best be sought in an XSL transformation rendering a bilingual HTML dictionary page.

Conclusions
In this paper, we have presented the existing Giella XML structure used in our MediaWiki based online dictionary. In addition, we have elaborated a way of converting from this XML formalism to the standardized TEI XML. This conversion is provided as an export functionality in our system.

Both the Giella XML and TEI have their own strengths and weaknesses. Supporting both of these formalisms makes it possible for us to combine the best form the both worlds. The Giella XML continues to be the primary import/export formalism for our synchronized MediaWiki-XML dictionary system because of its simplicity and integration with the Giella infrastructure. TEI is introduced as an additional export format for third parties to use the dictionary data in a standardized format.

References


