# Digitizing print dictionaries using TEI: The Abaev Dictionary Project

**Oleg Belyaev**
Lomonosov Moscow State University
Institute of Linguistics RAS
belyaev@ossetic-studies.org

**Irina Khomchenkova**
Vinogradov Russian Language Institute RAS
Lomonosov Moscow State University
irina.khomchenkova@yandex.ru

**Julia Sinitsyna**
Lomonosov Moscow State University
jv.sinitsyna@yandex.ru

**Vadim Dyachkov**
Institute of Linguistics RAS
hyppocentaurus@mail.ru

## Abstract

We present the results of a year-long effort to create an electronic version of V. I. Abaev's Historical-etymological dictionary of Ossetic. The aim of the project is two-fold: first, to create an English translation of the dictionary; second, to provide it (in both its Russian and English version) with a semantic markup that would make it searchable across multiple types of data and accessible for machine-based processing. Volume 1, whose prelimiary version was completed in 2020, used the TshwaneLex (TLex) platform, which is perfectly adequate for dictionaries with a low to medium level of complexity, and which allows for almost WYSIWYG formatting and simple export into a publishable format. However, due to a number of limitations of TLex, it was necessary to transition to a more flexible and more powerful format. We settled on the Text Encoding Initiative — an XML-based format for the computational representation of published texts, used in a number of digital humanities projects. Using TEI also allowed the project to transition from the proprietary, closed system of TLex to the full range of tools available for XML and related technologies. We discuss the challenges that are faced by such large-scale dictionary projects, and the practices that we have adopted in order to avoid common pitfalls.

## 1 Introduction

Digital lexicography is currently experiencing rapid development. With the transition to computerized publishing, most dictionaries are from the start conceived of as structured databases, with the print version being only one medium of many — and not a primary one at that. This, in most cases, presupposes a structure of lexical entries that is considerably different from that of earlier print dictionaries, where automatic processing was not an issue and the data were structured so as to be accessible in printed form. Major continuing publications (such as, for example, the Oxford English Dictionary (OED, 2021)) have already made the transition to digital formats. However, this is mainly true for large languages, where dictionaries are regularly published by stable teams having reliable financial support from state research institutions or private companies. For smaller languages, especially for minority languages, many dictionaries still only remain available in print (at best, scanned) form, with no possibility of automatic digitization due to the complexity of their structure and the inherent irregularity of their practical decisions (entry structure, choice of typefaces, etc.). Even when new dictionaries are published by local research teams, they are often prepared for typesetting as monolithic word-processor documents, making them largely equivalent to traditional print dictionaries prepared from card-catalogues — searchable by text, but without any semantic markup or more complex query mechanisms. This situation severely biases the range of lexicographical data available to researchers working on individual languages and in lexical typology — even when the dictionaries exist and are of a considerably high quality, they are virtually unavailable for automatic query and analysis.

This paper describes an attempt to fill this gap for Ossetic — an Iranian language spoken in the Caucasus by approximately 500 000 people. Ossetic is relatively well-documented lexicographically: bilingual (Abaev, 1970; Kasaev and Guriev, 1993; Takazov, 2003) and monolingual (Gæbæraty et al., 1999) dictionaries exist for both major dialects (Iron and Digor), and due to the effort of Ossetic language enthusiasts these have been converted into the ABBYY Lingvo format and an online searchable database (Iriston.com, 2004), which, while not ideal for research purposes and having some limitations, may at least be queried by headword.

However, the main lexicographic resource for Ossetic is still Vasily Ivanovič Abaev's fundamental, four-volume *Historical-etymological dictionary of Ossetic* (Abaev, 1958–1989) (henceforth AbD). This dictionary is not only one of the best etymological dictionaries available for any Iranian language (Zgusta, 1991), but also a very detailed descriptive, bilingual (Ossetic-Russian) dictionary — with the quality of definitions and the number of illustrative examples far surpassing that of all other Ossetic dictionaries. This dictionary still lacks a digital version, for obvious reasons: the structure of entries is complex and not trivial to capture in a standard dictionary format; the etymologies include examples from many different languages with diverse scripts that cannot be reliably OCR'd; manual verification should be undertaken. A further problem is that AbD is not available in English, making it unaccessible to scholars who do not fluently read Russian. By both digitizing and translating AbD, one would automatically provide a solid basis to further digital lexicographic work on Ossetic while also providing scholars with an English-based lexicographic resource for this language.

Therefore, with the encouragement of the Moscow Ossetian Fraternity (whose help and support we gratefully acknowledge), in the end of 2020 we began preliminary work on the project of both translating and digitizing AbD (including the Russian version, which should in any case be available as a benchmark against which uncertain parts of the translation can be verified). By the end of 2020, a first draft of the translation and database was prepared, published in a small number of copies in book form (Èto Kavkaz, 2020). This paper documents our experience with this project while highlighting the advantages and drawbacks of different approaches, and attempting to establish a best practice that could hopefully be used This was preceded by a preliminary analysis of the structure of Abaev's lexical entries, described in section 2. Section 3 describes the choice of TshwaneLex (TLex) (Joffe et al., 2021) as the software platform and the general structure as it was implemented by the end of 2020, and the disadvantages of TLex for a dictionary with a structure like AbD's. Finally, in section 4 we describe the transition to the Text Encoding Initiative (TEI) (The TEI Consortium, 2021) framework and the corresponding workflow, which solves most of the problems that we had with TLex and can serve as a useful foundation for further work on similar lexicographic projects — in particular for Uralic languages, given that a large number of similar legacy dictionaries are available for many of these languages, and Ossetic itself (unlike most other Iranian languages) is typologically similar to Uralic.

## 2 The structure of a dictionary entry in AbD

The overall structure of a mid-sized AbD lexical entry (that includes all the core elements but lacks additional complexities) can be illustrated by the lexeme *ad* 'taste', shown in Figure 1.

The entry can be subdivided into several clearly distinguished elements:

1. The **headword**, with a possible dialectal Digor form (separated from the main word by a vertical line).

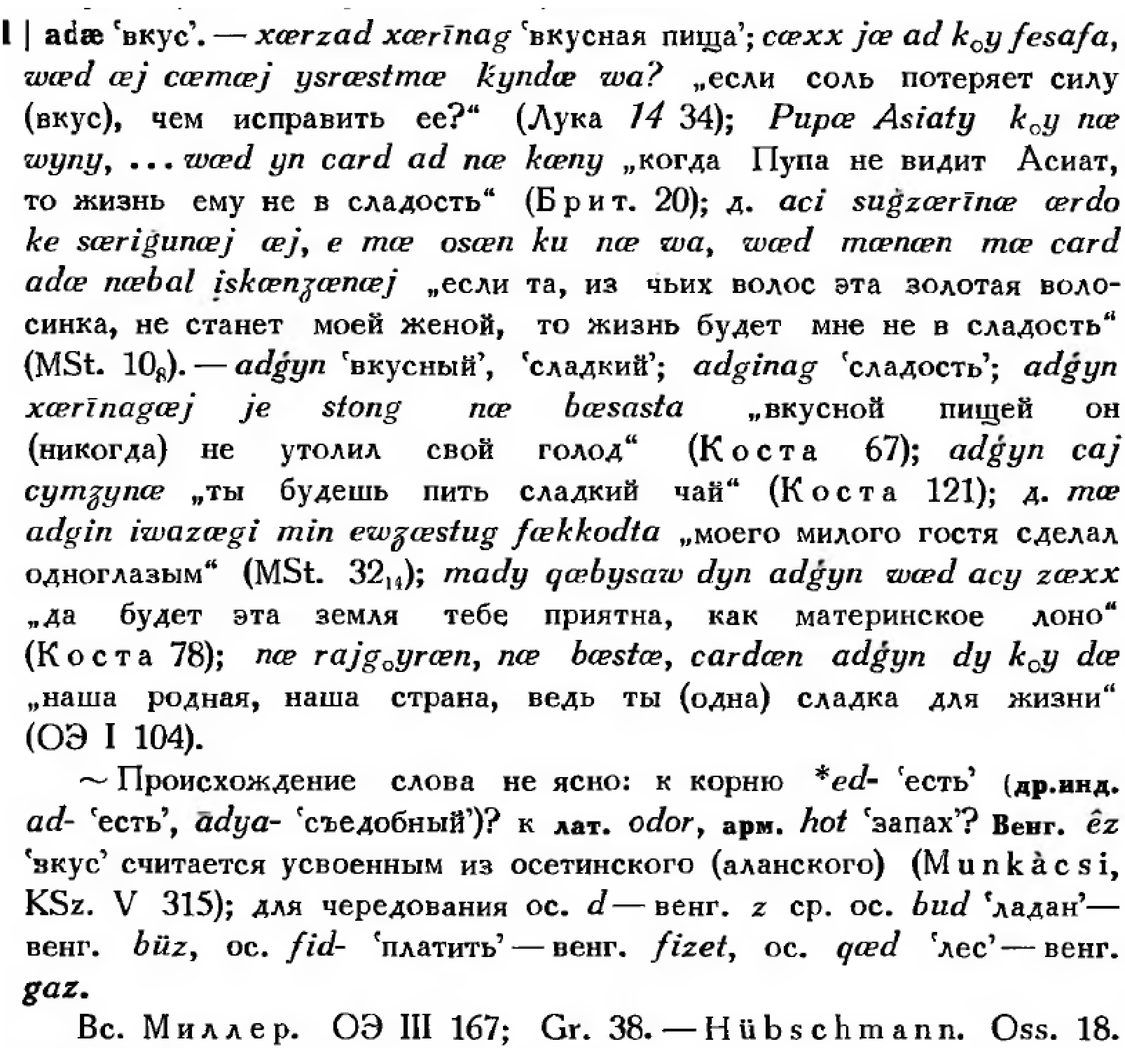
2. One or more **senses**, which consist of, most frequently, of short glosses in quotation marks, with possible additional comments.

3. An optional set of one or more **subentries** (idiomatic expressions or derivates from the headword), separated from each other by commas or semicolons; each subentry is a "mini-entry" in its own right, which may include several senses and its own examples.

4. One or more **groups of examples**; the group itself is separated by the surrounding content (including other example groups) by a dash, and examples are separated from each other by semicolons. The logic that stands behind using several groups of examples, rather than putting all exampels in one group, is in the general case not discernable. Sometimes both senses and example groups are numbered, in which case the group correspond to senses with the same numerical index.

5. An optional additional set of **subentries.**

6. A possible additional **example group** following the second set of subentries; only occurs



I | **adæ** ‘вкус’. — *xærzad xærīnag* ‘вкусная пища’; *cæxx jæ ad* $k_oy$ *fesafa, wæd æj cæmæj ysræstmæ kyndæ wa?* „если соль потеряет силу (вкус), чем исправить ее?“ (Лука *14* 34); *Pupæ Asiaty* $k_oy$ *næ wyny, . . . wæd yn card ad næ kæny* „когда Пупа не видит Асиат, то жизнь ему не в сладость“ (Брит. 20); д. *aci suǵzærīnæ ærdo ke særiǵunæj æj, e mæ osæn ku næ wa, wæd mænæn mæ card adæ næbal ịskænʒænæj* „если та, из чьих волос эта золотая волосинка, не станет моей женой, то жизнь будет мне не в сладость“ (MSt. $10_8$). — *adǵyn* ‘вкусный’, ‘сладкий’; *adginag* ‘сладость’; *adǵyn xærīnagæj je stong næ bæsasta* „вкусной пищей он (никогда) не утолил свой голод“ (Коста 67); *adǵyn caj cymʒynæ* „ты будешь пить сладкий чай“ (Коста 121); д. *mæ adgin iwazægi min ewʒæstug fækkodta* „моего милого гостя сделал одноглазым“ (MSt. $32_{14}$); *mady qæbysaw dyn adǵyn wæd acy zæxx* „да будет эта земля тебе приятна, как материнское лоно“ (Коста 78); *næ rajg*$_o$*yræn, næ bæstæ, cardæn adǵyn dy* $k_oy$ *dæ* „наша родная, наша страна, ведь ты (одна) сладка для жизни“ (ОЭ I 104).

~ Происхождение слова не ясно: к корню **ed-* ‘есть’ (**др.инд.** *ad-* ‘есть’, *ādya-* ‘съедобный’)? к **лат.** *odor*, **арм.** *hot* ‘запах’? **Венг.** *êz* ‘вкус’ считается усвоенным из осетинского (аланского) (Munkàcsi, KSz. V 315); для чередования ос. *d* — венг. *z* ср. ос. *bud* ‘ладан’— венг. *büz*, ос. *fid-* ‘платить’ — венг. *fizet*, ос. *qæd* ‘лес’ — венг. *gaz*.

Вс. Миллер. ОЭ III 167; Gr. 38. — Hübschmann. Oss. 18.

Figure 1: AbD entry for *ad* ‘taste’.

when the first block of example group(s) is also present.

7. The etymology, preceded by the tilde sign, which is essentially rich text which includes citations of forms from Ossetic and other languages (with the abbreviated language name typeset in bold) and bibliographic references.

This overall structure is of course a simplification: deviations from it are found in the dictionary, which is rather natural considering that the lexical entries were compiled by hand. However, in general, apart from the etymology, it is clear that the structure is relatively rigid so that it can be captured by a dictionary platform that allows custom data structures.

## 3 The TLex implementation

There are many lexicographic tools for linguists available today; the most popular ones are SIL Toolbox (SIL International, 2010) and Lexique Pro (SIL International, 2009), based on the Standard Format (SFM); and a more complex system implemented in SIL FieldWorks Language Explorer, or FLEx (SIL International, 2021). All these tools, while powerful and user-friendly, are aimed at field linguists documenting previously undescribed languages, and are ill-suited for a dictionary with such a non-standard structure as AbD. The standard tool for etymological dictionaries, StarLing (Starostin and Starostin, 2003), while powerful, is not suited for our purposes: it is rather deterministic, with the main aim being to capture exact etymological relationships, while AbD is in many cases ambiguous as to the exact etymology. Making an exact choice for an etymon is already an analytic decision that is beyond the scope of a digitization / translation project. Furthermore, StarLing provides little in terms of semantic markup. Therefore, we decided to choose another tool, also popular among digital lexicographers: TshwaneLex, or TLex (Joffe et al., 2021). This platform has been successfully used for numerous dictionary projects, notably the Beserman Udmurt dictionary, which has a complex structure comparable to that of AbD (Serdobolskaya et al., 2021). It is essentially a frontend to a highly customizable XML data structure. In particular, it is possible to define not only additional fields (as in FLEx), but a system of nested elements; importantly, the elements may contain mixed content with tags and PCDATA — this is essential for markup in the etymology to work correctly, as, of course, no rigid structure can account for the free-form text in Abaev's etymological descriptions. Accordingly, TLEx was used to implement a general dictionary structure that mimics the structure of Abaev's entries:

**LemmaSign** as an attribute (according to TLex usage), with optional LemmaVariant (for comma-separated orthographic / phonetic variants), Participle (verbs are quoted with participle forms, which are generally irregular) and DigorForm (for Digor dialectal forms, if they differ from Iron).

**PreSubentryGroup (0+)** a group of one or more subentries that precedes the first group of examples.

**ExampleGroup (0+)** the first set of example blocks;

**PostSubentryGroup (0+)** the second block of subentries;

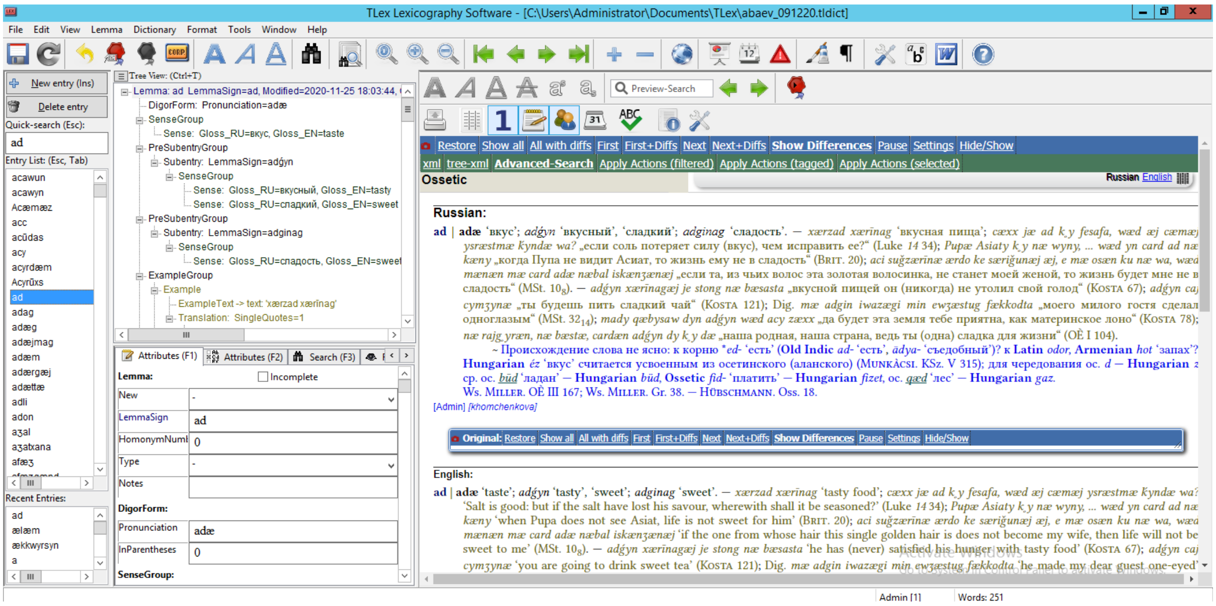
**ExampleGroup (0+)** the second set of example blocks;

**Etymology** with mixed content.

The structure of the same lexical entry *ad* ‘taste’ in TLex is shown in Figure 2.

The structure approximates AbD's structure relatively well and was used to successfully finish the translation of Vol. 1 of the dictionary. However, even from this general description of the structure, some problems are immediately apparent. For example, the difference between PreSubentryGroup and PostSubentryGroup seems to be completely artificial: these elements have exactly the same internal structure and display style. Following the logic of XML markup, they should definitely be assigned to the same element type.

The reason for this representation is the way TLex handles the order of elements: Unlike plain XML, where document order is always relevant, TLex ignores the order of elements in the XML source, only the structure is taken into account. This provides the advantage of being able to freely reorder elements using the built-in styling system. But the disadvantage is that it is practically impossible to differentiate between two or more elements that stand in different positions.



Quick-search (Esc): ad

Entry List: acawun, acawyn, Acæmæz, acc, acūdas, acy, acyrdæm, Acyrūxs, ad, adag, adæg, adæjmag, adæm, adærgæj, adættæ, adli, adon, aʒal, aʒatxana, afæʒ

Recent Entries: ad, ælæm, ækkwyrsyn, a

- Lemma: ad LemmaSign=ad, Modified=2020-11-25 18:03:44,
  - DigorForm: Pronunciation=adæ
  - SenseGroup
    - Sense: Gloss_RU=вкус, Gloss_EN=taste
  - PreSubentryGroup
    - Subentry: LemmaSign=adǵyn
      - SenseGroup
        - Sense: Gloss_RU=вкусный, Gloss_EN=tasty
        - Sense: Gloss_RU=сладкий, Gloss_EN=sweet
  - PreSubentryGroup
    - Subentry: LemmaSign=adginag
      - SenseGroup
        - Sense: Gloss_RU=сладость, Gloss_EN=sweet
  - ExampleGroup
    - Example
      - ExampleText -> text: 'xærzad xærīnag'
      - Translation: SingleQuotes=1

**Lemma:** ☐ Incomplete

| | |
|---|---|
| New | - |
| LemmaSign | ad |
| HomonymNumb | 0 |
| Type | - |
| Notes | |

**DigorForm:**

| | |
|---|---|
| Pronunciation | adæ |
| InParentheses | 0 |

**SenseGroup:**

**Ossetic**

**Russian:**

**ad** | **adæ** 'вкус'; *adǵyn* 'вкусный', 'сладкий'; *adginag* 'сладость'. — *xærzad xærīnag* 'вкусная пицца'; *cæxx jæ ad k̥y fesafa, wæd æj cæmæj ysræstmæ ǩyndæ wa?* „если соль потеряет силу (вкус), чем исправить ее?" (Luke *14* 34); *Pupæ Asiaty k̥y næ wyny, ... wæd yn card ad næ kæny* „когда Пупа не видит Асиат, то жизнь ему не в сладость" (Brit. 20); *aci suğzærīnæ ærdo ke særiğunæj æj, e mæ osæn ku næ wa, wæd mænæn mæ card adæ næbal iskænʒænæj* „если та, из чьих волос эта золотая волосинка, не станет моей женой, то жизнь будет мне не в сладость" (MSt. $10_8$). — *adǵyn xærīnagæj je stong næ bæsasta* „вкусной пищей он (никогда) не утолил свой голод" (Kosta 67); *adǵyn caj cymʒynæ* „ты будешь пить сладкий чай" (Kosta 121); Dig. *mæ adgin iwazægi min ewʒæstug fækkodta* „моего милого гостя сделал одноглазым" (MSt. $32_{14}$); *mady qæbysaw dyn adǵyn wæd acy zæxx* „да будет эта земля тебе приятна, как материнское лоно" (Kosta 78); *næ rajg̥yræn, næ bæstæ, cardæn adǵyn dy k̥y dæ* „наша родная, наша страна, ведь ты (одна) сладка для жизни" (OÈ I 104).

~ Происхождение слова не ясно: к корню **ed-* 'есть' (**Old Indic** *ad-* 'есть', *ādya-* 'съедобный')? к **Latin** *odor*, **Armenian** *hot* 'запах'? **Hungarian** *éz* 'вкус' считается усвоенным из осетинского (аланского) (Munkàcsi. KSz. V 315); для чередования ос. *d* — **Hungarian** *z* ср. ос. *būd* 'ладан' — **Hungarian** *büd*, **Ossetic** *fid-* 'платить' — **Hungarian** *fizet*, ос. *qæd* 'лес' — **Hungarian** *gaz*. Ws. Miller. OÈ III 167; Ws. Miller. Gr. 38. — Hübschmann. Oss. 18.

[Admin] *[khomchenkova]*

**English:**

**ad** | **adæ** 'taste'; *adǵyn* 'tasty', 'sweet'; *adginag* 'sweet'. — *xærzad xærīnag* 'tasty food'; *cæxx jæ ad k̥y fesafa, wæd æj cæmæj ysræstmæ ǩyndæ wa?* 'Salt is good: but if the salt have lost his savour, wherewith shall it be seasoned?' (Luke *14* 34); *Pupæ Asiaty k̥y næ wyny, ... wæd yn card ad næ kæny* 'when Pupa does not see Asiat, life is not sweet for him' (Brit. 20); *aci suğzærīnæ ærdo ke særiğunæj æj, e mæ osæn ku næ wa, wæd mænæn mæ card adæ næbal iskænʒænæj* 'if the one from whose hair this single golden hair is does not become my wife, then life will not be sweet to me' (MSt. $10_8$). — *adǵyn xærīnagæj je stong næ bæsasta* 'he has (never) satisfied his hunger with tasty food' (Kosta 67); *adǵyn caj cymʒynæ* 'you are going to drink sweet tea' (Kosta 121); Dig. *mæ adgin iwazægi min ewʒæstug fækkodta* 'he made my dear guest one-eyed'

Figure 2: The TLex representation of *ad* 'taste'

This leads to another artificial solution: the splitting of <Comment> elements into <PreComment> (before parent element) and <PostComment> (after parent element). But even this sometimes leads to absurd situation. For example, in our model, example texts and translations were originally represented by the attributes @text and @tr(_ru,en). A PreComment would then precede the example and a PostComment would follow the translation. Some examples, however, have a comment that stands between the example and the translation:

```
<PreComment> @text <MidComment>??
@tr <PostComment>
```

Clearly, a proliferation of <MidComment>-like elements is undesirable, because the range of possible comment positions can never be fully accounted for. The eventual solution was to repesent the example text and translation by elements (<ExampleText>, <Translation>), not attributes — which is, in fact, the natural way for XML, but not for TLex, which heavily favours attribute values and where adding new nested elements is a cumbersome process that is prone to error.

Another problem is the handling of styles. Surprisingly for an XML-based system (where CSS is normally available), TLex has a rather simplistic style system that cannot account for the element or attribute's context in any way. As seen in the above example, AbD uses punctuation patterns that are by themselves rather regular, but difficult for human annotators to consistently handle without error. It is therefore desireable to insert such regular punctuation automatically. In TLex, this can be done only by scripts written in an internal lua-based language. For example, the following code inserts a space before a <Source> (reference to an example source) that follows a <PostComment> element.

```
local prev = gCurrentNode:GetPrevious();
if prev ~= nil then
    if prev:GetElementTypeID() == 10079
    then
        gCurrentStyle:SetBeforeG(" ");
    end
end
```

The same functionality is easily captured in CSS by a single line:

```
  PostComment + Source::before {
content:  ''  ''}
```

The scripts are unnecessarily complex, written in a poorly documented language, and difficult to maintain; they may be adequate for comparatively minor dynamic styling, but as the project proceeded, it became clear that a large number of them is required. This made the dynamic punctuation practically unmanageable and difficult to debug.

A definite advantage of the TLex approach is support for controlled vocabularies, which here are called attribute lists (i.e. lists of possible values for certain attributes). However, this is not without a caveat: when server-based collaborative editing is used, any change to these attribute lists requires locking the whole database while making sure that all users have saved their data and logged out. This means that such trivial changes require an incomparable amount of effort, which complicates and slows down work on the dictionary.

To be sure the chief problem with using TLex for the AbD project is not that this is a bad piece of software — in fact, it is one of the best, if not the best, “off-the-shelf” dictionary creation tools currently available on the market. However, TLex’s use of XML is more suitable for relatively flat database structures where most of the information is stored in attributes. The use of mixed data and nested tags is complex and is not something TLex has been designed for. It is an adequate tool for new dictionary projects that follow a more modern, sense-based structure, or for digitization projects that also overhaul the structure of the original. When the aim is to represent the original as faithfully as possible, TLex is not the right tool for the job.

## 4 The TEI approach

When Volume 1 was finished, work began on converting the dictionary format to Text Encoding Initiative (TEI) Guidelines [REF], which define a set of tags and constraints for representation of texts in digital form. An immediate advantage of TEI compared to TLex is that, unless new tags are defined (which is seldom needed, because TEI is a very detailed standard) or existing tags abused, each element has a well-described semantics that is immediately accessible to any external observer, due to the structure being associated with the TEI namespace. TEI also represents displayed content primarily in elements rather than attributes (consistent with XML practice, which is, after all, a *markup* language) and is fully compatible with mixed-content elements. Thus, the example above is represented in TEI as follows:

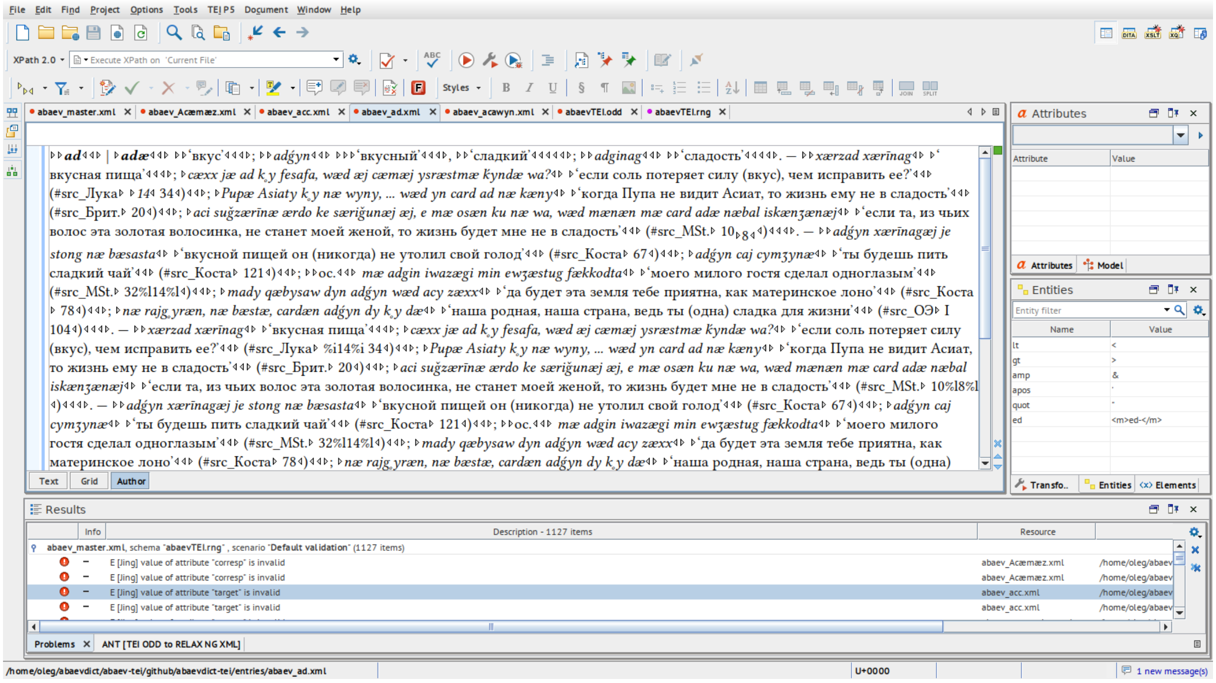
```
<cit type="example">
    <note type="comment">
        …(precomment)…
    </note>
    <quote>
        …(ex. text)…
    </quote>
    <note type="comment">
        …("midcomment")…
    </note>
    <cit type="translation"
        xml:lang=''ru''>
        …(translation)…
    </cit>
    <note type="comment">
        …("postcomment")…
    </note>
</cit>
```

Note the use of standard IETF BCP 47 (Network Working Group, 2009) language tags — this also allows interoperability and is implemented not only for English and Russian, but also for the Ossetic dialects and all languages cited in etymologies. The specific language strings can then be generate “on-the-fly” when the dictionary is converted (via XSLT or a similar transformation) into a publishable document.

An important feature of TEI is that it can be customize so that only the subset of all tags and attributes is selected that is actually required for a given project. This is done via files of a format called ODD (One Document Does (All)); our TEI customization is freely available in a GitHub repository: https://github.com/abaevdict/tei-abaev. This customization, of course, still remains rather redundant, allowing more than actually occurs; it could be constrained to resemble something like the rigid TLex schema above, but this is not required and in fact harmful, because further entries may include additional elements that have not been envisaged from prior experience (this being, after all, a legacy print dictionary).

The complex nested structures illustrated above can be edited in a user-friendly manner in modern XML editors such as Oxygen [REF], which we chose for this project. The editor natively supports TEI and allows “Author Mode” editing, which, styled with appropriate CSS, becomes almost a WYSIWYG model (see Figure 3). This significantly simplified work for the annotators, compared to TLex, where results are displayed in real-time, but the attributes and text values themselves have to be edited in a separate part of the screen.

The Oxygen customization, especially its CSS styles, are available on GitHub: https://github.com/abaevdict/abaev-tei-oxygen. The dictionary itself is split into multiple files, one file for each entry (generated from TLex using an XSLT transformation); the files are included in a single master file via XInclude. All dictionary data is also in a GitHub repo: https://github.com/abaevdict/abaevdict-tei. Collaborative editing can be done via standard



**ad** | **adæ** 'вкус'; *adǵyn* 'вкусный', 'сладкий'; *adginag* 'сладость'. — *xærzad xærīnag* 'вкусная пища'; *cæxx jæ ad k̥y fesafa, wæd æj cæmæj ysræstmæ kyndæ wa?* 'если соль потеряет силу (вкус), чем исправить ее?' (#src_Лука 14 34); *Pupæ Asiaty k̥y næ wyny, ... wæd yn card ad næ kæny* 'когда Пупа не видит Асиат, то жизнь ему не в сладость' (#src_Брит. 20); *aci suǧzærīnæ ærdo ke særiǧunæj æj, e mæ osæn ku næ wa, wæd mænæn mæ card adæ næbal iskænʒænæj* 'если та, из чьих волос эта золотая волосинка, не станет моей женой, то жизнь будет мне не в сладость' (#src_MSt. 10 8). — *adǵyn xærīnagæj je stong næ bæsasta* 'вкусной пищей он (никогда) не утолил свой голод' (#src_Коста 67); *adǵyn caj cymʒynæ* 'ты будешь пить сладкий чай' (#src_Коста 121); ос. *mæ adgin iwazægi min ewʒæstug fækkodta* 'моего милого гостя сделал одноглазым' (#src_MSt. 32%l14%l); *mady qæbysaw dyn adǵyn wæd acy zæxx* 'да будет эта земля тебе приятна, как материнское лоно' (#src_Коста 78); *næ rajg̥yræn, næ bæstæ, cardæn adǵyn dy k̥y dæ* 'наша родная, наша страна, ведь ты (одна) сладка для жизни' (#src_ОЭ I 104). — *xærzad xærīnag* 'вкусная пища'; *cæxx jæ ad k̥y fesafa, wæd æj cæmæj ysræstmæ kyndæ wa?* 'если соль потеряет силу (вкус), чем исправить ее?' (#src_Лука %i14%i 34); *Pupæ Asiaty k̥y næ wyny, ... wæd yn card ad næ kæny* 'когда Пупа не видит Асиат, то жизнь ему не в сладость' (#src_Брит. 20); *aci suǧzærīnæ ærdo ke særiǧunæj æj, e mæ osæn ku næ wa, wæd mænæn mæ card adæ næbal iskænʒænæj* 'если та, из чьих волос эта золотая волосинка, не станет моей женой, то жизнь будет мне не в сладость' (#src_MSt. 10%l8%l). — *adǵyn xærīnagæj je stong næ bæsasta* 'вкусной пищей он (никогда) не утолил свой голод' (#src_Коста 67); *adǵyn caj cymʒynæ* 'ты будешь пить сладкий чай' (#src_Коста 121); ос. *mæ adgin iwazægi min ewʒæstug fækkodta* 'моего милого гостя сделал одноглазым' (#src_MSt. 32%l14%l); *mady qæbysaw dyn adǵyn wæd acy zæxx* 'да будет эта земля тебе приятна, как материнское лоно' (#src_Коста 78); *næ rajg̥yræn, næ bæstæ, cardæn adǵyn dy k̥y dæ* 'наша родная, наша страна, ведь ты (одна)

| Name | Value |
|---|---|
| lt | < |
| gt | > |
| amp | & |
| apos | ' |
| quot | " |
| ed | <m>ed-</m> |

| Info | Description - 1127 items | Resource | |
|---|---|---|---|
| | abaev_master.xml, schema "abaevTEI.rng", scenario "Default validation" (1127 items) | | |
| – | E [Jing] value of attribute "corresp" is invalid | abaev_Acæmæz.xml | /home/oleg/abaev |
| – | E [Jing] value of attribute "corresp" is invalid | abaev_Acæmæz.xml | /home/oleg/abaev |
| – | E [Jing] value of attribute "target" is invalid | abaev_acc.xml | /home/oleg/abaev |
| – | E [Jing] value of attribute "target" is invalid | abaev_acc.xml | /home/oleg/abaev |

/home/oleg/abaevdict/abaev-tei/github/abaevdict-tei/entries/abaev_ad.xml U+0000 1 new message(s)

Figure 3: The representation of *ad* 'taste' in Oxygen's Author mode

Git mechanisms, which is essentially error-proof, because each annotator keep a full local copy of the database on their machine. The use of GitHub also allows for undisruptively making modifications to the schema files: the annotators need only pull the relevant repositories, without the need to "lock" the database.

The elements corresponding to the TLex structure illustrated above are as follows:

**form** the head word (with the `@type = 'lemma'` property) or various variants (with `@type = 'dialectal'` or `@type = 'inflected'`, and various subtypes of inflected forms);

**sense** the sense information block that contains definitions or translations in Russian or English;

**re** corresponds to (Pre/Post)SubentryGroup; the xGroups are not actually needed because TEI XML is position-aware.

**cit** with `@type = 'exampleGroup'` is admittedly a slight deviation from TEI semantics, given that an example group is not an example itself. However, it is fairly close, because it may only containt `cit` elements which are examples.

**etym** the etymology block, with mixed content.

Thus, using TEI, the dictionary ends up with a structure that is more complex in some sense, but at the same time less rigid and having less limitations than the TLex model.

## 5 Conclusion

This paper describes the experience of our research group in an attempt to achieve a double aim: provide a translation of AbD and also digitize it, supplying it with semantic markup. Of course, this is not the first legacy dictionary project that utilizes TEI (Du Fresne Du Cange et al., 1883–1887; Littré, 1863–1873),[1] but the specific challenge is unique due to both its double aim and the complexity (and partial ambiguity) of AbD's structure. In the talk, we will discuss the dictionary structure, its implementation in TLex and TEI, and the corresponding problems in more detail, attempting to provide a set of best practices for digitizing traditional etymological dictionaries.

## References

Vasilij I. Abaev. 1958–1989. *Istoriko-ètimologičeskij slo-*

[1] There is even a modification of TEI for lexicography, called TEI-Lex-0 (DARIAH Working Group "Lexical Resources", 2020), but it is too rigid for the purposes of faithfully representing an older print dictionary. In particular, mixed content is discouraged, which, in the case of Abaev's etymologies, is impossible to follow.

*var' osetinskogo jazyka*. Nauka, Moscow, Leningrad. [Historical-etymological dictionary of Ossetic]. Vols. 1–4. In Russian.

Vasily I. Abaev. 1970. *Russko-osetinskij slovar'*. Nauka, Moscow. [Russian-Ossetic dictionary]. In Russian.

Charles Du Fresne Du Cange et al. 1883–1887. *Glossarium mediæ et infimæ latinitatis*. Niort. Online edition. accessed 07.03.2021.

DARIAH Working Group "Lexical Resources". 2020. *TEI Lex-0: A baseline encoding for lexicographic data*. Accessed 07.03.2021.

Èto Kavkaz. 2020. Istoriko-ètimologičeskij slovar' osetinskogo jazyka pereveli na anglijskij. [The Historical-Etymological dictionary of Ossetic has been translated into English] News item. Accessed 07.03.2021.

Nik'ala Gæbæraty, Tamerlan G˳yriaty, Nafi Ǯusojty, Šamil Ǯykkajty, and Xarum Taqazty. 1999. *Iron ævzaǯy æmbyryngænæn ʒyrdwat*. Tskhinval, Vladikavkaz, Vladikavkaz. [Explanatory dictionary of Ossetic]. In Ossetic.

Iriston.com. 2004. Slovari na iriston.com. [Dictionaries on IRISTON.COM] In Russian. Accessed 07.03.2021.

David Joffe et al. 2021. *TLex Lexicography*. Accessed 07.03.2021.

Alexander M. Kasaev and Tamerlan Aleksandrovič Guriev, editors. 1993. *Osetinsko-russkij slovar'*, 4th edition edition. Izdatel'stvo Severo-Osetinskogo instituta gumanitarnyx i social'nyx issledovanij, Vladikavkaz. [Ossetic-Russian dictionary.] About 28000 words. In Russian.

Émile Littré. 1863–1873. *Dictionnaire de la langue française*. Paris. Accessed 07.03.2021.

Network Working Group. 2009. *Tags for Identifying Languages (BCP 47)*. Accessed 07.03.2021.

OED. 2021. *The Oxford English Dictionary*. OUP, Oxford. Accessed 07.03.2021.

Natalia Serdobolskaya et al. 2021. Beserman-Russian dictionary. Accessed 07.03.2021.

SIL International. 2009. Lexique Pro. Accessed 07.03.2021.

SIL International. 2010. *Field Linguist's Toolbox*. Accessed 07.03.2021.

SIL International. 2021. *FieldWorks Language Explorer*. Accessed 07.03.2021.

Sergei A. Starostin and George S. Starostin. 2003. *The Tower of Babel: An international etymological database project*. Accessed 07.03.2021.

Fedar M. Takazov, editor. 2003. *Digorsko-russkij slovar'*. Vladikavkaz. [Digor-Russian dictionary]. In Russian.

The TEI Consortium. 2021. *P5: Guidelines for Electronic Text Encoding and Interchange*. Accessed 07.03.2021.

Ladislaw Zgusta. 1991. Typology of etymological dictionaries and V. I. Abaev's Ossetic Dictionary. pages 38–49.