

Электронная версия Историко-этимологического словаря осетинского языка В.И. Абаева и проблемы создания электронных этимологических словарей

О. И. Беляев^{1,2}, И. А. Хомченкова^{1,2,3}, Ю. В. Синицына², В. В. Дьячков²

¹МГУ имени М. В. Ломоносова,

²Институт языкознания РАН,

³Институт русского языка РАН

Наш проект

- Цель проекта – создание электронной двуязычной версии четырёхтомного *Историко-этимологического словаря* [Абаев 1958–1989], далее ИЭСОЯ.
 - семантическая разметка: возможность поиска по различным типам данных, создание интерактивных версий, автоматическая обработка;
 - перевод на английский язык.
- В докладе мы обсудим:
 - значимость данных осетинского языка для лингвистики и истории;
 - ИЭСОЯ, его ценность и уникальность;
 - ход нашего проекта, технические решения и результаты.
- Мы благодарны за поддержку РНФ, проект № 22-28-01639 «Создание двуязычной цифровой версии *Историко-этимологического словаря осетинского языка В.И. Абаева*», а также инициатору и вдохновителю проекта А. М. Торчинову и московской осетинской общине.

Зачем?

- Многие словари малых языков доступны только в печатной (в лучшем случае, в сканированной) форме. Оцифровать их автоматически невозможно из-за сложности структуры и нерегулярности практических решений (структура статьи, выбор шрифтов, нестандартная транскрипция и т.д.).
 - В таком виде они недоступны ни для поиска, ни для какой-либо автоматической обработки. Создание на их базе электронных ресурсов с гипертекстовыми функциями также невозможно.
 - Современные словари в подавляющем большинстве случаев перешли на цифровые форматы, где собственно печатный вариант является лишь одной из форм представления исходных данных (если он вообще издаётся).
- Мы хотим также иметь электронный вариант ИЭСОЯ.

Зачем?

- Осетинский (иранский, ок. 500 000 носителей) имеет довольно неплохую базу словарей [Абаев 1970; Касаев 1993; Таказов 2003; Агнаев 1999], которые энтузиасты конвертировали в электронный формат **ABBYY Lingvo** и сайт с поиском (<https://slovar.iriston.com>).
- Однако ИЭСОЯ остаётся самым подробным и точным источником, причём он является как одним из лучших этимологических словарей для иранских языков в целом [Zgusta 1991], так и очень хорошим описательным словарём.
- **Проблема:** сложная структура статьи; этимология включает примеры из множества различных языков с разными письменностями, которые не могут быть надёжно распознано → нужно вручную размечать, проверять и переводить данные.

Как?

- Том 1 (2020): платформа **TshwaneLex** (TLex)
 - + отлично подходит для словарей с уровнем сложности от низкого до среднего;
 - + позволяет задавать форматирование практически в формат WYSIWYG, простой экспорт в стандартный формат (.doc);
 - имеет ряд ограничений.
 - Потребовалось перейти на более гибкий формат.
- Том 2 (2022), другие тома (в процессе): **Text Encoding Initiative** (TEI)
 - формат на основе XML, специально разработанный для представления письменных / печатных текстов; стандарт в области цифровых гуманитарных наук (Digital humanities);
 - использование TEI также позволило перейти от закрытой (хотя и также основанной на XML) системы TLex к открытому формату и использованию стандартных технологических средств в сфере работы с XML.

Структура словарной статьи в ИЭСОЯ

- 1 **Заглавное слово**
- 2 **Значение** (одно или более)
- 3 **Подстатьи** (факультативно; идиоматические выражения или слова, образованные от заглавного); каждая подстатья является «мини-статьёй», может включать несколько значений, собственные примеры и т. д.
- 4 Группа **примеров** (одна или более)
- 5 **Подстатьи** (факультативный второй блок)
- 6 Группа **примеров** (факультативный второй блок)
- 7 **Этимология**

k'ori 'шар'; *zæxxy k'ori* 'земной шар'. — *k'orijy x_oyzæn tymbyl læppā*
„мальчик, круглый, как шар“ (С е к а 126).
~ Из араб., перс. *kura*, тюрк. *küre* 'шар'.

Структура словарной статьи в ИЭСОЯ

***ǵzælyn** (*yzǵælyn*): ***ǵzæld** | **æǵzælnun**: **æǵzald** 'сыпаться', 'осыпаться', 'капать'; **ǵzæl** 'то, что осыпалось, пролилось, капнуло', 'кроха', 'зерно', 'кровавый след'; д. *æǵzald* 'хлеб в зерне' (обычно кукуруза). — *æx-sæly yzǵæly* „можжевельник осыпается“ (Коста 125); *cæstysyg yzǵæly agmæ sūsægæj* „слезы текут украдкой в котел“ (Коста 44); д. *kudtæncæ madæltæ, næ xwærtæ, æǵzaldæj cæstitæn sæ don* „плакали матери, наши сестры, струилась влага из глаз“ (Bes. 71); д. *avaræn niǵzaldæj je sarst* „у здания осыпалась штукатурка“ (SD 295₃). — *xæzary cy mūr, zǵæl wyd, adæmæn sæ baxæryn kodta* „все крохи, какие еще были в доме, пошли на угощение народа“ (Сека 47); д. *zænxætæ æǵzæl dær næ niggælstǫncæ* „они не бросили в землю ни зернышка“ (SD 173₄); д. ... *næ xæzari ewnæg æǵzæl dær ævard ku næ wa, otæ* „... чтобы в нашем доме не было спрятано ни крохи“ (FS II 65); д. *æǵzælbæl cæwun rajdædta Sostan* „Сослан стал идти по кровавому следу“. — *fydyzǵæl* | *fidæǵzæl* 'мясо' („кроха мяса“). — д. *kæd æǵzaldi put radzænæ, kena ba dæ bæx fæstæmæ lasæ* „если дашь пуд кукурузы, а не то води свою лошадь обратно“ (SD 280₃).

~ Медиальное соответствие к *ǵzælyn* 'сыпать' (РОСл. 574).

Этимология

æ̆gzalyn (*yz̆gzalyn*): *æ̆gzald* | *æ̆gzalun*: *æ̆gzald* ‘сыпать’, ‘осыпать’, ‘ссыпать’; медиально (с ослаблением гласного) *gz̆æl̆yn* ‘сыпаться’, ‘осыпаться’, ‘капать’.

~ Вряд ли основательно сближение с др.инд. *gal-* ‘капать’, ‘падать’ (Вс. Миллер. Gr. 57); *gal-* — изолированный в арийских языках, специфический санскритский (в Риг-Веде отсутствует) глагол, непригодный для целей осетинской этимологии; ос. *l* в *gz̆alyn* не может быть отождествлен с др.инд. *l* [... es läge nahe anzunehmen, dass *l* im Ai. durchaus nur jüngere Entwicklung des *r* wäre und der indoiranische Rhotazismus alle ig. *l* beseitigt hätte, bevor wieder ein spezifisch ai. *l* neu aufkam“ (J. Wackernagel. Altindische Grammatik I 216 сл.)]. Более приемлемой кажется ранняя этимология того же Вс. Миллера (ОЭ II 59): к ав. *γ̆zar-*, *žgar-* ‘течь’, ‘литься’, каузативно — ‘заставлять течь’, ‘заливать’, др.инд. *k̆sar-*, *k̆sarati* ‘течь’, каузативно ‘лить’ (об отношении др.инд. *k̆s* к ав. *γ̆z* см.: Brugmann I 617). Расхождение значений ‘сыпать’ — ‘лить’ кажущееся: *æ̆gzald* ‘кровавый след, оставляемый раненым животным или человеком’ показывает, что *gz̆alyn*, *gz̆æl̆yn* применялись в прошлом не только к сыпучим телам, но и к жидкости (*æ̆gzald̆bæl̆ cæ̆wun̆ rajdæ̆dta Soslan* ‘Сослан пошел по кровавому следу’). Ср. также *cæ̆stysy̆g yz̆gæ̆ly* ‘слезы текут’ (Коста 44). Tomaschek (891) относит сюда же перс. *šārīdan* ‘течь’, нам. с. *pa-x̆čor-* ‘лить’, афг. *z̆γ̆al-*, *z̆γ̆astal* ‘бежать’, а также ос. *z̆goryn* ‘бежать’.

Вс. Миллер. ОЭ II 59.

Tlex

Общая информация

- Программный пакет для работы со словарями [Joffe et al. 2021]
- Ранее использовался в других проектах, например, в проекте словаря бесермянского удмуртского (<http://beserman.ru/>)
- Настраиваемая структура, основанная на XML

TLex

Структура

LemmaSign атрибут (в соответствии с конвенциями TLex), с факультативными атрибутами **LemmaVariant** (для орфографических / фонетических вариантов), **Participle** (с глаголами всегда цитируются формы причастий, обычно нерегулярные) и элементом **DigorForm** (для форм на дигорском диалекте, если отличаются от иронских);

Sense (1+) значение;

PreSubentryGroup (0+) группа подстатей перед первым блоком примеров;

ExampleGroup (0+) первый блок групп примеров;

PostSubentryGroup (0+) второй блок подстатей;

ExampleGroup (0+) второй блок групп примеров;

Etymology текстовое содержание, размеченное тэгами.

Интерфейс Tlex

The screenshot displays the Tlex Lexicography Software interface. The main window shows a dictionary entry for the Russian word "осыпаться" (to spill, to crumb). The entry is structured as follows:

- Lemma:** *gʒælyn, LemmaSign=*gʒælyn, Modified=2020-11-30 18:14:50, Created=2020-11-30 18:14:50
- Participle:** Pronunciation=*gʒæld
- LemmaVariant:** Pronunciation=yʒgælyn, InParentheses=1
- DigorForm:** Pronunciation=aʒgælyn
- Participle:** Pronunciation=aʒgæld
- SenseGroup:**
 - Sense:** Gloss_RU=осыпаться, Gloss_EN=to pour
 - Sense:** Gloss_RU=осыпаться, Gloss_EN=to fall
 - Sense:** Gloss_RU=канать, Gloss_EN=to drop
- PreSubentryGroup:**
 - Subentry:** LemmaSign=*gʒæln
 - SenseGroup:**
 - Sense:** Gloss_RU=то, что осыпалось, пролилось, капнуло, Gl
 - Sense:** Gloss_RU=кроха, Gloss_EN=crumb
 - Sense:** Gloss_RU=зерно, Gloss_EN=seed
 - Sense:** Gloss_RU=крошечный след, Gloss_EN=blood trace
 - PostComment:** CommentType=Semicolon before
 - Text_RU** → text: д. *aʒgældʲaʲ 'xleb s zerne' (обычно кукуруза)
 - Text_EN** → text: Digor *aʒgældʲaʲ 'grain' (usually corn/)
- ExampleGroup:**
 - Example:**
 - ExampleText** → text: 'aʒsæly yʒgæly'
 - Translation:**
 - Text_RU** → text: 'можевелиник осыпается'
 - Text_EN** → text: 'juniper's leaves are falling down'
 - Source:** Abbreviation=%skocra%5, Text=125
 - Example:**
 - ExampleText** → text: 'caestysy yʒgæly agmae sʲusægaej'
 - Translation:**
 - Text_RU** → text: 'слезы текут украдкой в котел'
 - Text_EN** → text: 'teardrops stealthily fall into the cauldron'
 - Source:** Abbreviation=%skocra%5, Text=44
 - Example:** Type=д.
 - ExampleText** → text: 'kudtaencae madæltæ, nae xwærtæ, aʒgældæj caestitaen sae don'
 - Translation:**
 - Text_RU** → text: 'плакали матери, наши сестры, струилась влага'
 - Text_EN** → text: 'our mothers, our sisters were crying, the moisture was flowing'

The main text area shows the entry in Russian and English. The Russian text is as follows:

Russian:

*gʒælyn (yʒgælyn) : *gʒæld | aʒgælyn : aʒgæld 'сыпаться', 'осыпаться', 'капать'; gʒæln 'то, что осыпалось, пролилось, капнуло', 'кроха', 'зерно', 'кровавый след'; д. aʒgæld 'хлеб в зерне' (обычно кукуруза).. — aʒsæly yʒgæly 'можевелиник осыпается' (Коста 125); caestysy yʒgæly agmae sʲusægaej 'слезы текут украдкой в котел' (Коста 44); д. kudtaencae madæltæ, nae xwærtæ, aʒgældæj caestitaen sae don 'плакали матери, наши сестры, струилась влага из глаз' (Бес. 71); д. avaræn niʒgældæj je sarst 'у здания осыпалась штукатурка' (SD 295). — хаэзэгу су мур, yʒgæly wyd, adætaen sae baxæryn kodta 'все крохи, какие еще были в доме, пошли на угощение народа' (Сека 47); д. zaenxaeae aʒgæld daer nae niʒgælstocae 'они не бросили в землю ни зернышка' (SD 173); д. ...nae хаэзэри ewnae aʒgæld daer avarad ku nae wa otæ '...чтобы в нашем доме не было спрятано ни крохи' (FS II 65); д. aʒgælbæln caewun rajdædta Soslan 'Сослан стал идти по кровавому следу'. — fʲidʲy-gæln | fʲidægʒæln 'мясо' ('кроха мяса'). — д. kaed aʒgældi put radzaenae, kena ba dae bæx faestæmae lasæ 'если дашь пуд кукурузы, а не то води свою лошадь обратно' (SD 280).

— Медальное соответствие к *gʒælyn 'сыпаться' (РОСЯ. 574).

[Admin] | khomchenkova]

The English text is as follows:

English:

*gʒælyn (yʒgælyn) : *gʒæld | aʒgælyn : aʒgæld 'to pour', 'to fall', 'to drop'; gʒæln 'thing that fell or spilled', 'crumb', 'seed', 'blood trace'; Digor aʒgæld 'grain' (usually corn). — aʒsæly yʒgæly 'juniper's leaves are falling down' (Коста 125); caestysy yʒgæly agmae sʲusægaej 'teardrops stealthily fall into the cauldron' (Коста 44); д. kudtaencae madæltæ, nae xwærtæ, aʒgældæj caestitaen sae don 'our mothers, our sisters were crying, the moisture from the eyes was flowing' (Бес. 71); д. avaræn niʒgældæj je sarst 'The plaster of a building was peeled off.' (SD 295). — хаэзэгу су мур,

TLex

Проблемы

- **pre & postsubentry group, pre & postcomment...** → умножение сущностей

почему? TLex игнорирует порядок элементов в исходном файле, учитывается только структура (кто в кого вложен). Отдельный набор настроек «стиля» определяет порядок. Но если есть несколько элементов одного типа в смешанном порядке (например, SubentryGroup > ExampleGroup > SubentryGroup), то невозможно задать их взаимный порядок.

- **Midcomment?** Изначально тексты примеров и переводы обозначались атрибутами text и tr(_ru,en). Тогда элемент PreComment предшествует примеру, а PostComment – следует. Но в некоторых примерах комментариев стоит между текстом примера и его переводом.

Tlex

Проблемы

- Редактирование ведётся онлайн на сервере; если подключение теряется, значительная часть данных может быть утрачена.
- Наборы стандартных терминов (например, названия языков) можно изменить, только заблокировав всю базу.
- Автоматическая пунктуация реализована на основе lua, API плохо документировано и результат невозможно экспортировать в другие приложения. Например:

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

TEI

Общее описание

- Универсальный набор элементов и ограничений для представления электронных версий текстов [Tei Consortium 2021].
- Содержит модули для различных проектов в области цифровых гуманитарных наук, включая словари [TEI Lex-0 2020]
- Можно настроить под конкретный проект, выбрав тот набор тегов, который нужен, и добавив свои при необходимости.
- Мы максимально ограничили TEI для ИЭСОЯ, добавив только три новых элемента для упрощения аннотации:

`abv:example` = `<cit type="example">`

`abv:exampleGrp` = `<cit type="exampleGrp">`

`abv:tr` = `<cit type="translation">`

TEI

Пример

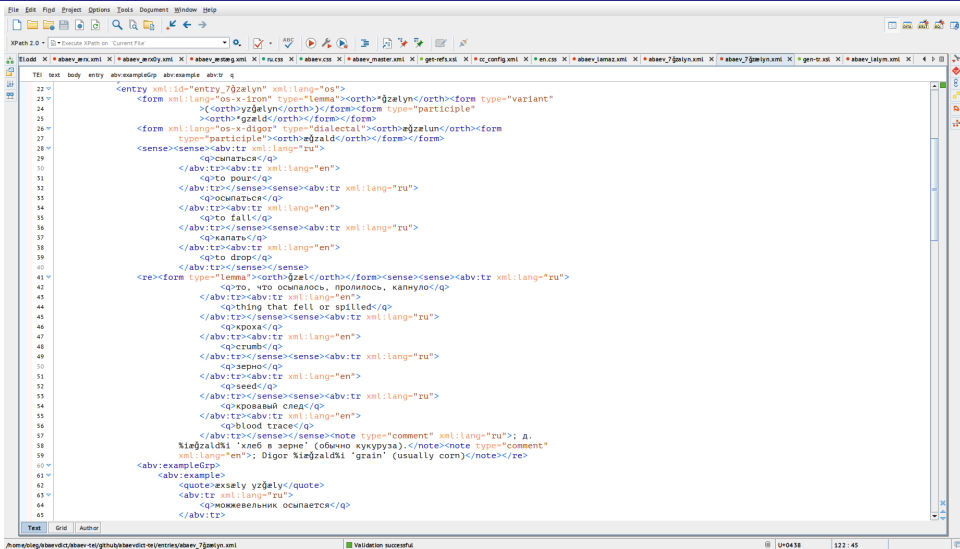
```
<abv:example>
  <note type="comment">
    ...(precomment)...
  </note>
  <quote>
    ...(ex. text)...
  </quote>
  <note type="comment">
    ...("midcomment")...
  </note>
```

```
<abv:tr>
  ...(translation)...
</abv:tr>
<note type="comment">
  ...("postcomment")...
</note>
</abv:example>
```

TEI

Редактирование

- Разработан специально настроенный framework для Oxygen XML Editor (<https://www.oxygenxml.com/>).
- Это позволяет автоматически вставлять элементы и атрибуты, а также редактировать практически в формате WYSIWYG.
- Визуализация через CSS, что гораздо проще чем скрипты в TLex:
 - `note[type="comment"] + bibl::before { content: " " }`
- Для публикации можно использовать трансформацию XSL в LaTeX.
- Контроль версий при помощи Git (см.: <https://github.com/abaevdict/>).

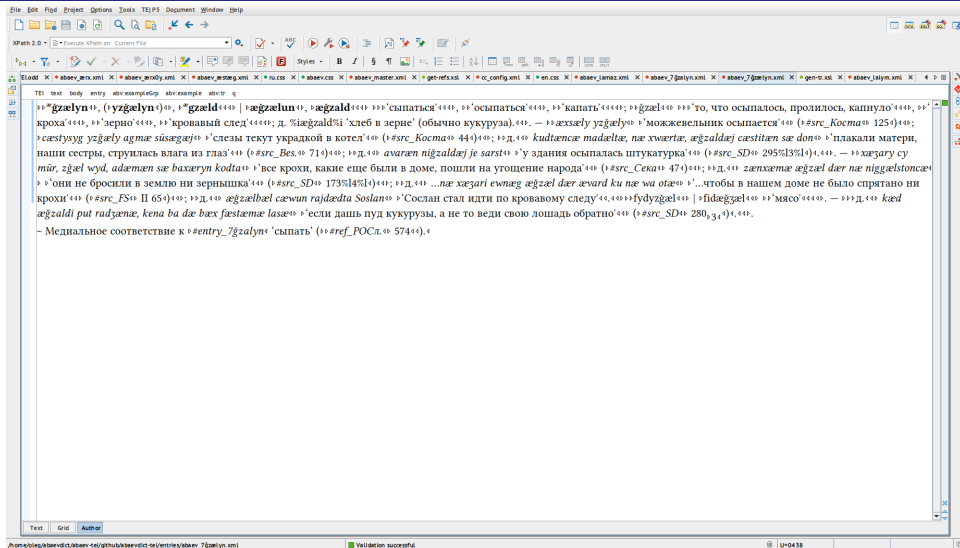


```
File Edit Find Project Options Tools Document Window Help
XPath 2.0 - Execute XPath on 'Current File'
Tlx X abev_entr.xml X abev_entrDy.xml X abev_entrDy.xml X ru.css X abev.css X abev_master.xml X get-refs.xml X cr_config.xml X en.css X abev_lamz.xml X abev_7qzaln.xml X abev_7qzaln.xml X gen-tr.xml X abev_lalym.xml X
TEI text body entry abv-exampleGrp abv-example abv-tr q
22 <entry xml:id="entry_7qzaln" xml:lang="os">
23   <form xml:lang="os-x-iron" type="lemma"><orth>*qzaln</orth><form type="variant"
24     ><orth>yzqzaln</orth></form><form type="participle"
25     ><orth>*qzald</orth></form></form>
26   <form xml:lang="os-x-digor" type="dialectal"><orth>qzaln</orth><form
27     type="participle"><orth>qzald</orth></form></form>
28   <sense><sense><abv:tr xml:lang="ru">
29     <q>сына́ться</q>
30   </abv:tr><abv:tr xml:lang="en">
31     <q>to pour</q>
32   </abv:tr></sense><sense><abv:tr xml:lang="ru">
33     <q>осына́ться</q>
34   </abv:tr><abv:tr xml:lang="en">
35     <q>to fall</q>
36   </abv:tr></sense><sense><abv:tr xml:lang="ru">
37     <q>кана́ть</q>
38   </abv:tr><abv:tr xml:lang="en">
39     <q>to drop</q>
40   </abv:tr></sense></sense>
41   <re><form type="lemma"><orth>qzal</orth></form><sense><sense><abv:tr xml:lang="ru">
42     <q>то, что осыналось, пролилось, кануло</q>
43   </abv:tr><abv:tr xml:lang="en">
44     <q>thing that fell or spilled</q>
45   </abv:tr></sense><sense><abv:tr xml:lang="ru">
46     <q>кпо́ха</q>
47   </abv:tr><abv:tr xml:lang="en">
48     <q>crumb</q>
49   </abv:tr></sense><sense><abv:tr xml:lang="ru">
50     <q>зе́рно</q>
51   </abv:tr><abv:tr xml:lang="en">
52     <q>seed</q>
53   </abv:tr></sense><sense><abv:tr xml:lang="ru">
54     <q>крова́вый сле́д</q>
55   </abv:tr><abv:tr xml:lang="en">
56     <q>blood trace</q>
57   </abv:tr></sense></sense><note type="comment" xml:lang="ru">; д.
58   %qzald%и 'хлеб в зерне' (обычно кукуруза).</note><note type="comment"
59   xml:lang="en">; Digor %qzald% 'grain' (usually corn)</note></re>
60   <abv:exampleGrp>
61     <abv:example>
62       <quote>exsely yzqzaln</quote>
63       <abv:tr xml:lang="ru">
64         <q>може́веньник осына́ется</q>
65       </abv:tr>
66     </abv:example>
67   </abv:exampleGrp>
68 </entry>
```

Text Grid Author

Validation successful

U+0438 122 / 45



Выводы

- При оцифровке или переводе печатных словарей нужно использовать семантическую разметку.
- TLex удобен, но в большей степени подходит для словарей-баз данных, но не для этимологических словарей с гораздо более свободной структурой и большим количеством цитированных слов и примеров.
- Для ИЭСОЯ в долгосрочной перспективе предпочтителен TEI:
 - хорошо документированные стандарты (XML, XSL, CSS);
 - стандартный набор элементов с ясной семантикой;
 - более гибкая настройка под конкретные нужды;
 - лучше подходит для этимологических словарей.