Working with lexical data: Experiences with Uralic languages

Course: NLP for Endangered Languages of the Amazon. From a Uralic perspective. Lecture 3.

Jack Rueter, Niko Partanen, Mika Hämäläinen & Khalid Alnajjar

Dictionaries

Arranging a lexicon is a central task in language documentation

Dictionaries also have a large role in language planning and teaching

Two distinct, but connected, work phases:

- Dictionary creation (i.e. for endangered languages)
- Dictionary maintenance (adding new terms, editing meanings)

Especially in multilingual dictionaries the relationships are important

- Derivational relations, variants, synonyms, antonyms, etymologies etc.

Dictionary editing

There are several commonly used tools and databases:

- FLEx
- TLex
- Lexique Pro
- Lexonomy

- In rule based natural language processing we often use so called LEXC files or various XML standards
- None of these solutions address very well the connections between multilingual dictionaries, and the need for multiple uses for the same data

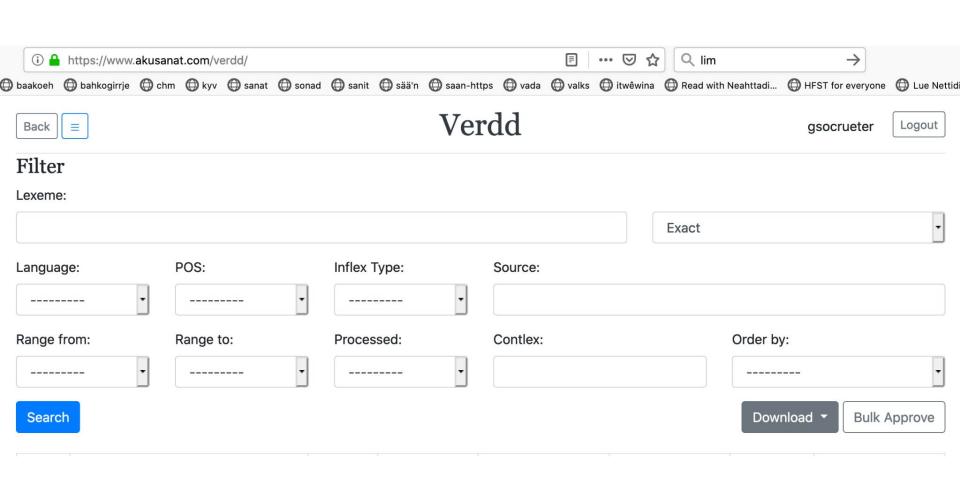
Ve'rdd – an open-source dictionary editing platform

The goal has been to combine various dictionaries so that entries can be linked Not tied to a specific format:

- Uses an internal database, but imports and exports different structures
- Looking for a way to keep ahead of XML and JSON

Write a word or annotation ONCE

Still in experimental phase, but used with several research groups in Finland



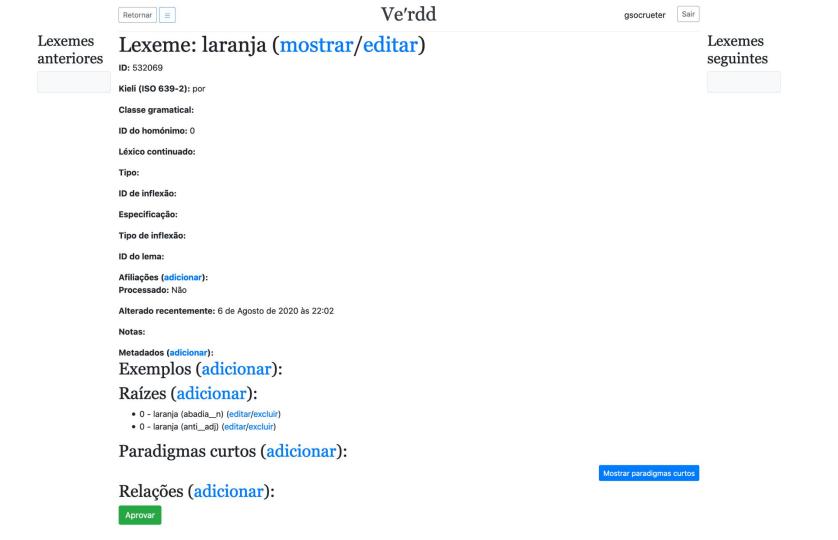


Total: 1





Total: 1



Relações (adicionar):

Aprovar

Retornar	Ve'rdd
Adicionar relação:	
Desde: laranja	
Até	
Type*	
Tradução	
Notes	
Notas	



Adicionar relação:

Desde: laranja

Até

Type* √Tradução Etimologia Palavra composta Derivação Variante Fraseologia Hyponymy Hyperonymy Holonymy Meronymy Co-hyponym Synonym Antonym Outro

Retornar

Ve':

Adicionar relação:

appelsiininkokoinen (A) - 0 fin

Desde: laranja

Até appelsiini appelsiini (N) - 0 krl appelsiini (N) - 0 fin appelsiini- (A) - 0 fin appelsiinijuoma (N) - 0 fin appelsiinilehto (N) - 0 fin appelsiinilikööri (N) - 0 fin appelsiinilimonadi (N) - 0 fin appelsiinimarmeladi (N) - 0 fin appelsiinimehu (N) - 0 fin appelsiinimehutiiviste (N) - 0 fin appelsiinin (A) - 0 fin appelsiininkeltainen (A) - 0 fin



Adicionar relação:

Desde: laranja

Até

appelsiini (N) - 0 fin

Type*

Tradução

Notes

Notas

Salvar

Retornar

Desde

ID: 532069

Kieli (ISO 639-2): por

Classe gramatical:

ID do homónimo: 0

Léxico continuado:

ID de inflexão:

Especificação:

Tipo de inflexão:

Processado: Não

Afiliações (adicionar):

Alterado recentemente: 6 de Agosto de 2020 às 22:02

ID do lema:

Notas:

Tipo:

Lexeme: laranja

(mostrar/editar)

Lexeme: appelsiini (mostrar/editar)

Até ID: 30542 Kieli (ISO 639-2): fin Classe gramatical: N ID do homónimo: 0 Léxico continuado: Tipo: ID de inflexão: Especificação: Tipo de inflexão: ID do lema: Afiliações (adicionar):

• Akusanat: Fin:appelsiini (editar/excluir)

Alterado recentemente: 6 de Agosto de 2020 às 18:16

Processado: Não

Metadados (adicionar):

Metada

MOLAS:

Relação:

Kieli (ISO 639-2):

Tipo: Tradução

Processado: Não

Notas:

Alterado recentemente: 29 de Junho de 2021 às 14:03

Fontes

• adicionar

Exemplos

adicionar

Metadados

• adicionar

Aprovar

Lexemes anteriores	Lexer 1D: 532069	ne: lara	anja (<mark>m</mark> o	strar	'editar)				
lampeiro	Kieli (ISO 63	9-2): por							
lamuriador	Classe gram	atical:							
lanceador	ID do homór	nimo: 0							
lanceolado	Léxico conti	nuado:							
lancetado	Tipo:								
lancetável	ID de inflexâ	io:							
lanchado	Especificaçã	io:							
lancinante	Tipo de infle	xão:							
landês	ID do lema:								
landeiro	Afiliações (a								
langoroso	Processado	: Não							
lanoso	Alterado rec	entemente: 6 de	e Agosto de 2020 à	s 22:02					
lanuginoso	Notas:								
lançado	Metadados Exemn	(adicionar): olos (adic	ionar).						
lançador	_	(adicion							
lançável		nja (abadia_n) (e							
laparotomizável		nja (abadia_n) (ed nja (anti_adj) (ed							
lapidado	Paradi	gmas cur	rtos (adicio	onar):					
lapidificável	_ ~~~~	0	(44101	, ,				Mostra	r paradigmas curtos
lapidoso	Relaçõ	es (adicio	onar):						
lapidário	ID	Desde	Até	Tipo	Fontes	Exemplos	Metadados	Notas	Ações
lapidável	10	Desue	Are		rontes	LACITIPIOS	Metauauus	NULAS	Vônca
lapidífico	1139182	(por) laranja	(fin) appelsiini	Tradução	adicionar	adicionar	adicionar		mostrareditar
laqueado									• excluir
laqueável									

Ve'rdd Retornar Adicionar relação: Desde: laranja Até fruta desfrutado () - 0 por desfrutador () - 0 por desfrutar () - 0 por desfrutação () - 0 por fruta (N) - 0 por frutado () - 0 por

Retornar

Adicionar relação:

Desde: laranja

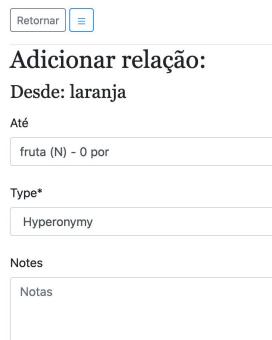
Até

Type*

Hyperonymy

Notes

Notas



Notas

fronteira
frota
frouxidade
frugalidade
frugalismo

fruição

frumento

frumentação

frustração

Paradigmas curtos (adicionar):

Relações (adicionar):

0 - II uta (abadia_II) (edital/excluii)

ID	Desde	Até	Tipo	Fontes	E
1094577	(apu) txipukury	(por) fruta	Tradução	• adicionar	
1139192	(por) laranja	(por) fruta	Hyperonymy	• adicionar	

Aprovar

ID	Desde	Até	Tipo	Fontes	Exemplos	Metadados	Notas	Ações
1094577	(apu) txipukury	(por) fruta	Tradução	• adicionar	• adicionar	• adicionar		mostrareditarexcluir

Desde	Até
Lexeme: txipukury	Lexeme: fruta
(mostrar/editar)	(mostrar/editar)
ID: 674676	ID: 554277
Kieli (ISO 639-2): apu	Kieli (ISO 639-2): por
Classe gramatical: N	Classe gramatical: N
ID do homónimo: 0	ID do homónimo: 0
Léxico continuado: N_SUFFIXES_01_Msc_re2/ry	Léxico continuado:
Tipo:	Tipo:
ID de inflexão:	ID de inflexão:
Especificação:	Especificação:
Tipo de inflexão: X	Tipo de inflexão:
ID do lema:	ID do lema:
Afiliações (adicionar): Processado: Não	Afiliações (adicionar): Processado: Não
Alterado recentemente: 6 de Agosto de 2020 às 23:03	Alterado recentemente: 4 de Maio de 2021 às 19:20
Notas:	Notas:
Metadados (adicionar):	Metadados (adicionar):



Filtrar

Lexeme:				É idêntico +
Língua:	Classe gramatical:	Tipo de inflexão:	Fonte:	
Desde:	Até:	Processado:	Léxico continuado:	Ordem:
Pesquisar	_	_		Baixar ▼ Aprovar tudo

ID	Lexeme	Classe gramatical	Léxico continuado	Tipo de inflexão	Língua	Notas	Ações
674622	hãkiti	N	N_SUFFIXES_01_Msc	X	apu		mostrareditarexcluir
674623	herẽkatxi	N	N_SUFFIXES_01_Msc_0/txi	X	apu	herēka >> herēkatxi	mostrareditarexcluir

Retornar	≡
----------	---

Ve'rdd

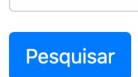
Filtrar

1	ex	10	m	0	
ш	-	$^{\prime}$		C.	

Língua:

•

Desde:	



Classe gramatical:



Até:

Processado:

Tipo de inflexão:



Fonte:

(
l.			





Retornar	≡	Ve'rdd				ć	gsocruet	
Filtrar								
					É idê	ntico		
Língua:		Classe gramatical:	Tipo de inflexão:	Fonte:				
apu	_	N						
Desde:		Até:	Processado:	Léxico continuado:		Orde	m:	
	•							
Pesquisar							Baixar 🕶	Apr
ID	Lexeme	Classe gramatical	Léxico continuado	Tipo de inflexão	Língua	Notas	CSV LEXC	
674622	hãkiti	N	N_SUFFIXES_01_Msc	X	apu			

```
hãkiti+v1+N:hãkiti N_SUFFIXES_01_Msc "" ;
hãkiti+v2+N:ãkiti N_SUFFIXES_01_Msc ""
herēkatxi+v1+N:herēka N SUFFIXES 01 Msc 0/txi "" ;
herēkatxi+v2+N:erēka N_SUFFIXES_01_Msc_0/txi "" ;
herēkatxi+v3+N:arēka N_SUFFIXES_01_Msc_0/txi ""
amaru+N:amaru N_SUFFIXES_01_Fem-ru_te/0 "" ;
amary+N:amary N_SUFFIXES_01_Msc-ry_te/0 "" ;
kanawa+N:kanawa N_SUFFIXES_01_Msc_te/0 "" ;
iuwata+N:iuwata N_SUFFIXES_01_Fem_ne/0 "" ;
namatxi+N:nama N_SUFFIXES_01_Msc_0/txi ""
ukypitetxi+N:ukypite N_SUFFIXES_01_Msc_0/Ntxi "";
kiritãtxi+N:kirita N_SUFFIXES_01_Msc_0/Ntxi "" ;
xinhîtxi+N:xinhi N_SUFFIXES_01_Msc_0/Ntxi "" ;
wakũtxi+N:waku N SUFFIXES 01 Msc 0/Ntxi "" ;
pitxîtxi+N:pitxi N_SUFFIXES_01_Msc_0/Ntxi "" ;
sukūtxi+Hom1+N:suku N_SUFFIXES_01_Msc_0/Ntxi
```

674626,apu,kanawa,N,N_SUFFIXES_01_Msc_te/0,X 674627,apu,iuwata,N,N_SUFFIXES_01_Fem_ne/0,X 674628,apu,namatxi,N,N_SUFFIXES_01_Msc_0/txi,X 674629,apu,ukypitetxi,N,N_SUFFIXES_01_Msc_0/Ntxi,X 674630,apu,kiritatxi,N,N_SUFFIXES_01_Msc_0/Ntxi,X 674631,apu,xinhîtxi,N,N_SUFFIXES_01_Msc_0/Ntxi,X

674632,apu,wakütxi,N,N_SUFFIXES_01_Msc_0/Ntxi,X

674634,apu,sukütxi,N,N_SUFFIXES_01_Msc_0/Ntxi,X

674633,apu,pitxîtxi,N,N_SUFFIXES_01_Msc_0/Ntxi,X

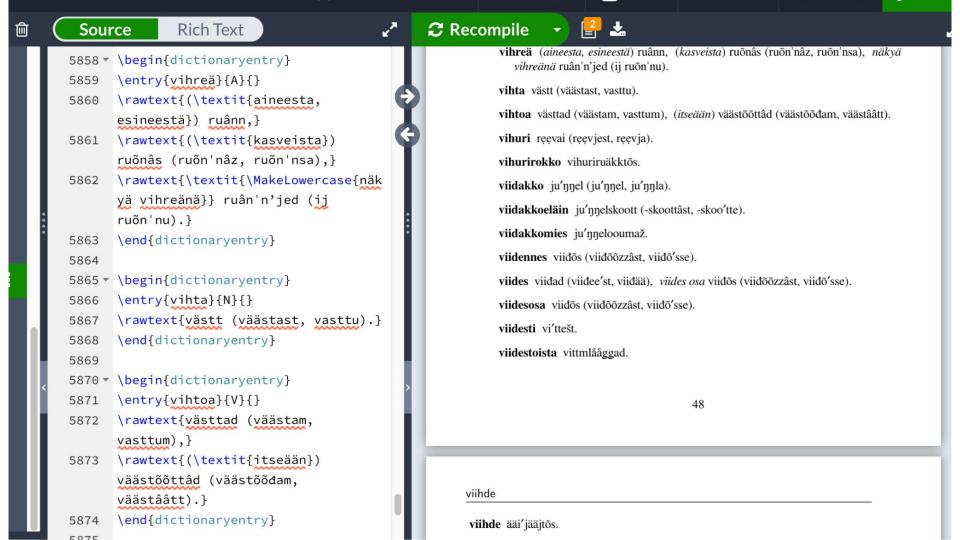
674623,apu,herekatxi,N,N_SUFFIXES_01_Msc_0/txi,X

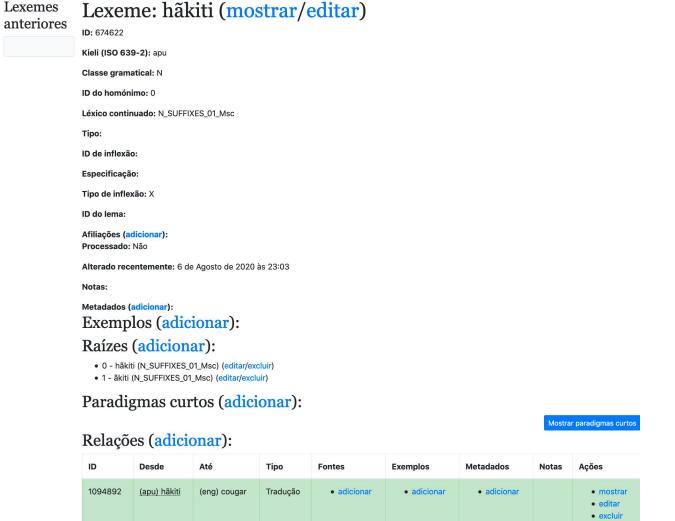
.624,apu,amaru,N,N_SUFFIXES_01_Fem-ru_te/0,X

674625, apu, amary, N, N_SUFFIXES_01_Msc-ry_te/0, X

ID, Language, Lexeme, POS, Contlex, Inflex Type

674622, apu, hãkiti, N, N_SUFFIXES_01_Msc, X





Lexemes

Notas:

Metadados (adicionar):

Exemplos (adicionar):

Raízes (adicionar):

- 0 hãkiti (N_SUFFIXES_01_Msc) (editar/excluir)
- 1 ãkiti (N_SUFFIXES_01_Msc) (editar/excluir)

Paradigmas curtos (adicionar):

Relações (adicionar):

ID	Desde	Até	Tipo
1094892	(apu) hãkiti	(eng) cougar	Tradução

Paradigmas curtos (adicionar):

ID	Descrição morfossintática	Forma da palavra
-	N+PI+Gen	appelsiinien
-	N+PI+Nom	appelsiinit
-	N+Sg+III	appelsiiniin

Mostrar paradigmas curtos

Ve'rdd & FLEx

FLEx has different export formats

For example, LIFT XML file could be imported

It's possible to imagine a workflow where FLEx lexicon is exported and synchronized regularly into Ve'rdd, where more complex relations are stored

We have done this once with dictionaries stored in MediaWiki

We would be happy to experiment with a new language that has FLEx corpus!