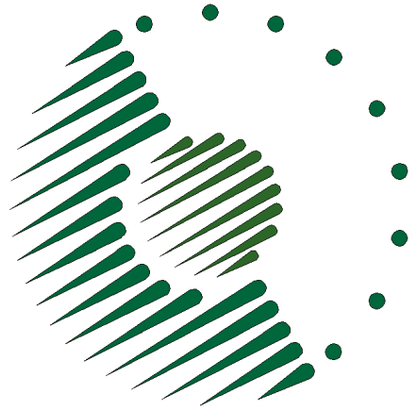
Panlingual Localization

Jonathan Pool
Utilika Foundation
pool@utilika.org



LISA @ Berkeley
5 August 2009

The Localization Industry Standards Association

Panlingual Localization

1. Goal
2. Strategy
3. Method
4. Demonstration
5. Challenges
6. Solutions
7. Discussion

1. Goal: Panlingual Localization

Make any information instantly and freely accessible in any language.

~~Google: 127~~

~~Unicode CLDR: 145~~

~~Universal Declaration of Human Rights: 360~~

~~Wikipedia: 235~~

Utilika Foundation

2. Strategy: Lemmatic Communication

Все счастливые семьи похожи друг на друга, каждая несчастливая семья несчастлива по-своему.

все	счастливый	семья	похожий
все	несчастливый	семья	особенный

tout	heureux	famille	similaire
tout	malheureux	famille	particulier

Les familles heureuses se ressemblent toutes; les familles malheureuses sont malheureuses chacune à leur façon.

3. Method: Mine Diverse Lexical Resources

Wiktionaries

VikiSözlük
Özgür Sözlük

Thesauri

γλώσσα για ειδικούς σκοπούς
MT (70.20)
Da: fagsprog
De: Fachsprache
En: language for special purposes
Es: lenguaje especializado
Fi: kieli tiettyihin tarkoituksiin
Fr: langage spécialisé
He: שפה למטרות מיוחדות
Hu: szaknyelv
It: lingua speciale
Nl: vaktaal
Sv: fackspråk
BT γλώσσες

Dictionaries

Locale Repositories

SE (English: Sweden)

An tSualainn	·ga ·
isveç	·az ·
İsveç	·tr ·
Iswidhan	·so ·
Rootsi	·et ·
Ruotsi	·fi ·
Ruotta	·se ·
Schweden	·de ·
Schweede	·gsw ·

Glossaries

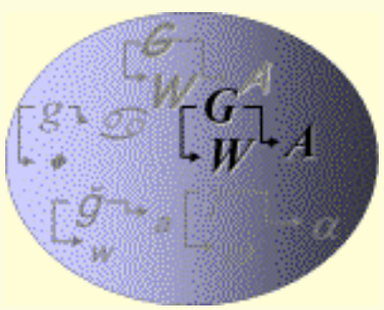
Arrest: ᑎᒍᖏᐅᓂᖅ: Tigujauniq: Arrestation

The act of placing a person in custody, according to law. The powers of ordinary citizens and peace officers to arrest a person are set out in the *Criminal Code*, 1996, Part XVI.

Arson: ᐃᑭᑎᕋᑎᓂᖅ: Ikitittiniq: Crime d'incendie

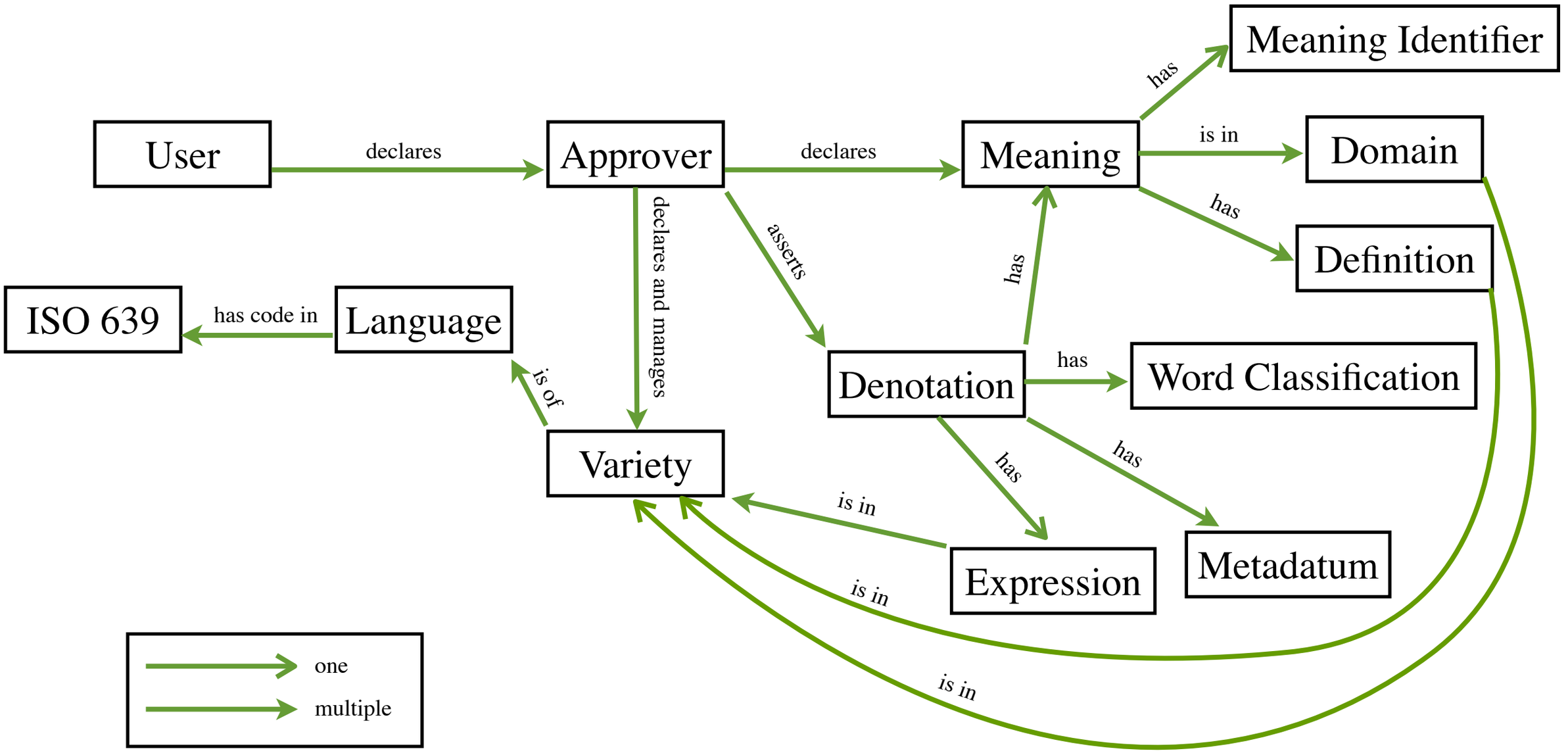
The crime of deliberately setting fire to property. *Criminal Code*, 1996, sections 433-436.

WordNets



Utilika Foundation

3. Method: Unify Heterogeneous Data



Utilika Foundation

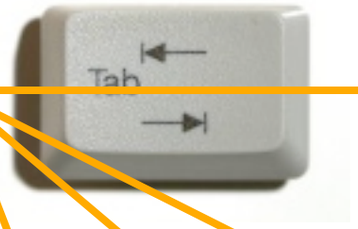
3. Method: Infer New Results

F 1

F 2

hnappur
Icelandic

3 D



klavo
Esperanto

stud
English

button
English

Taste
German

touche
French

tasto
Italian

tecla
Spanish

клавиша
Russian

key
English

4 A

5 A

6 M

7 M

8 M

9 M

10 M

11 M

12 M

13 M

14 M

15 U

16 U

17 A

الزرّ
Arabic

الزّرّ
Arabic

زر
Arabic

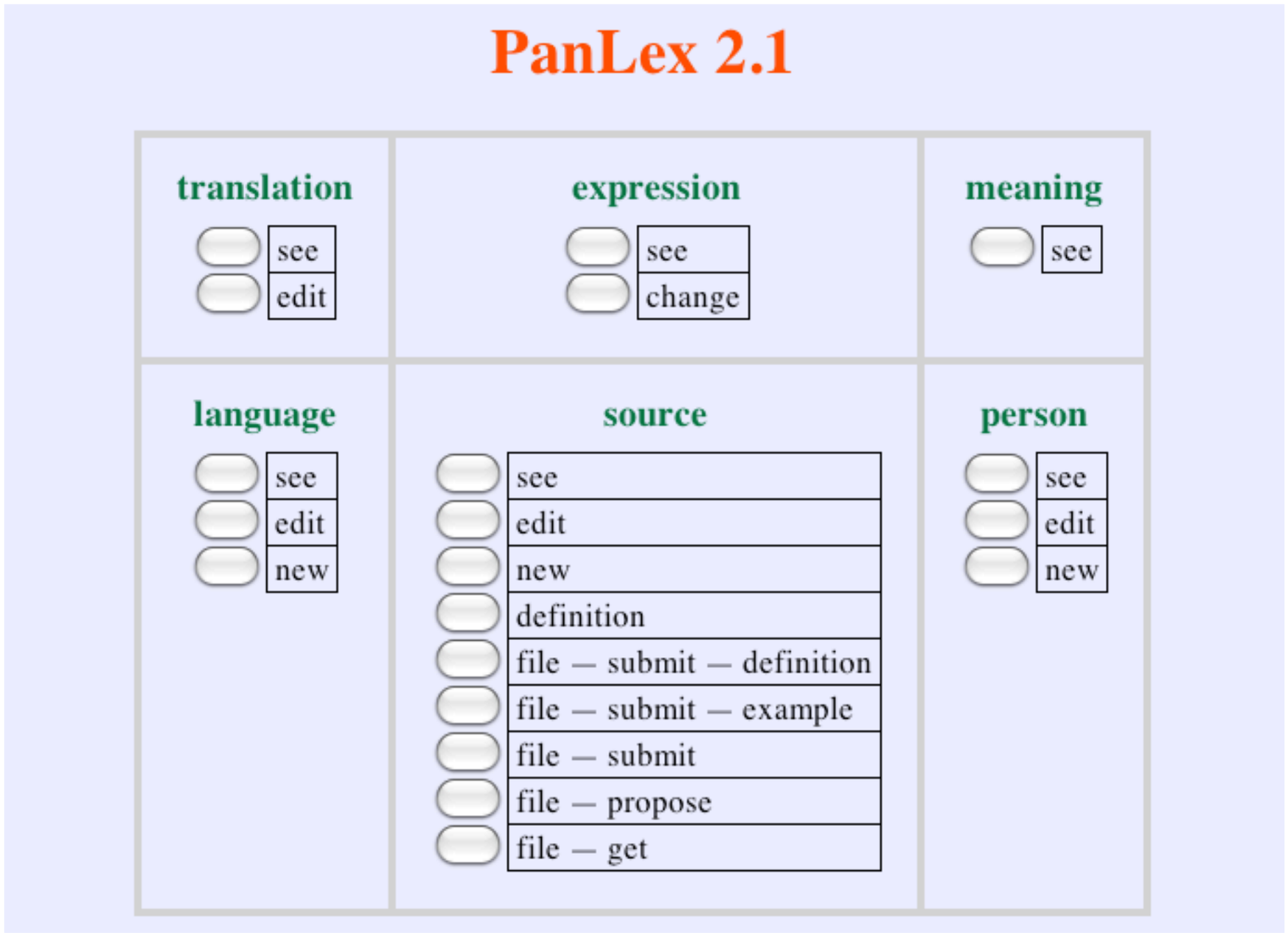
مفتاح
Arabic

المفتاح
Arabic

Utilika Foundation

4. Demonstration

Lemmatic UI



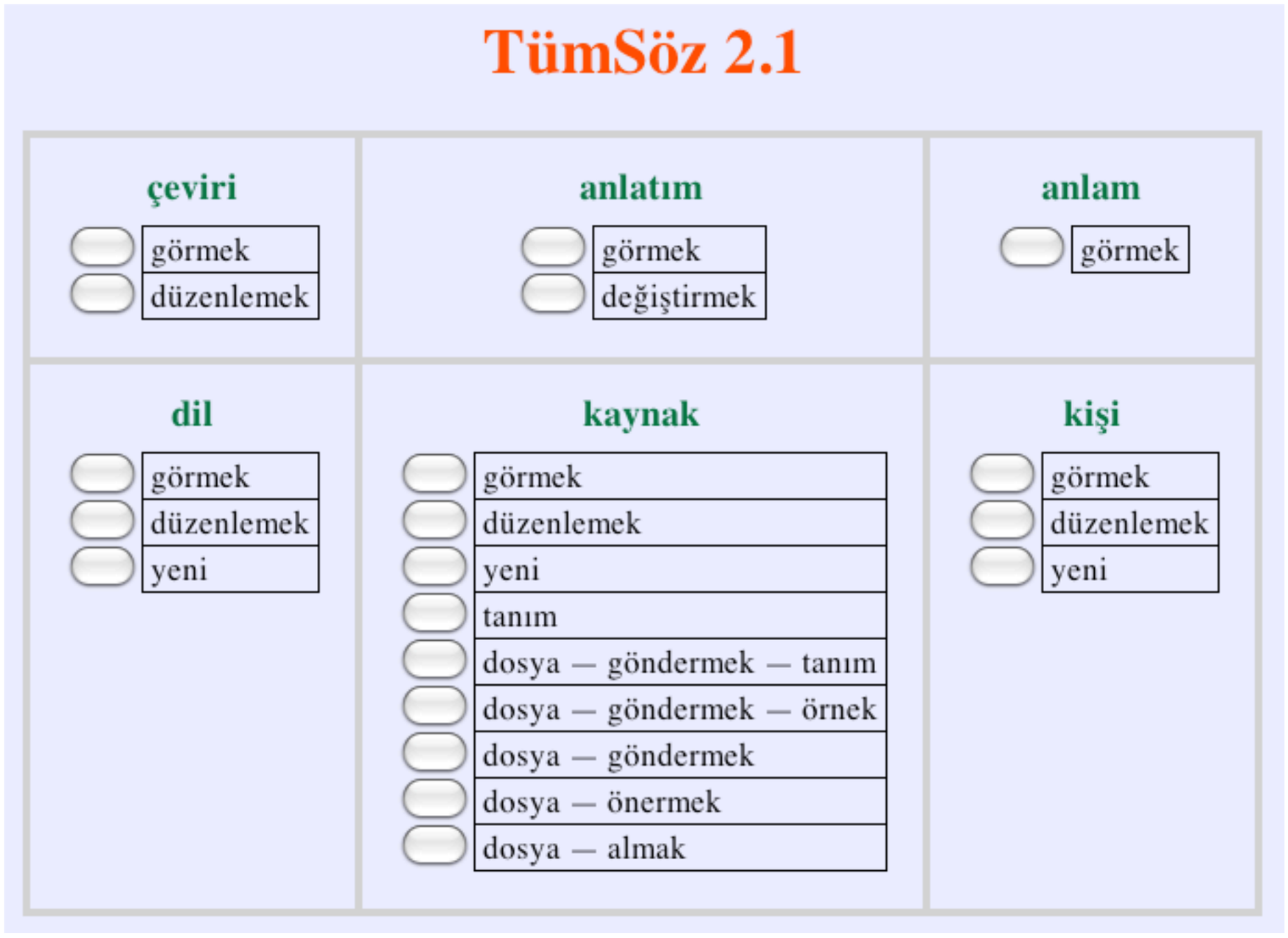
English

http://panlex.org/u

Utilika Foundation

4. Demonstration

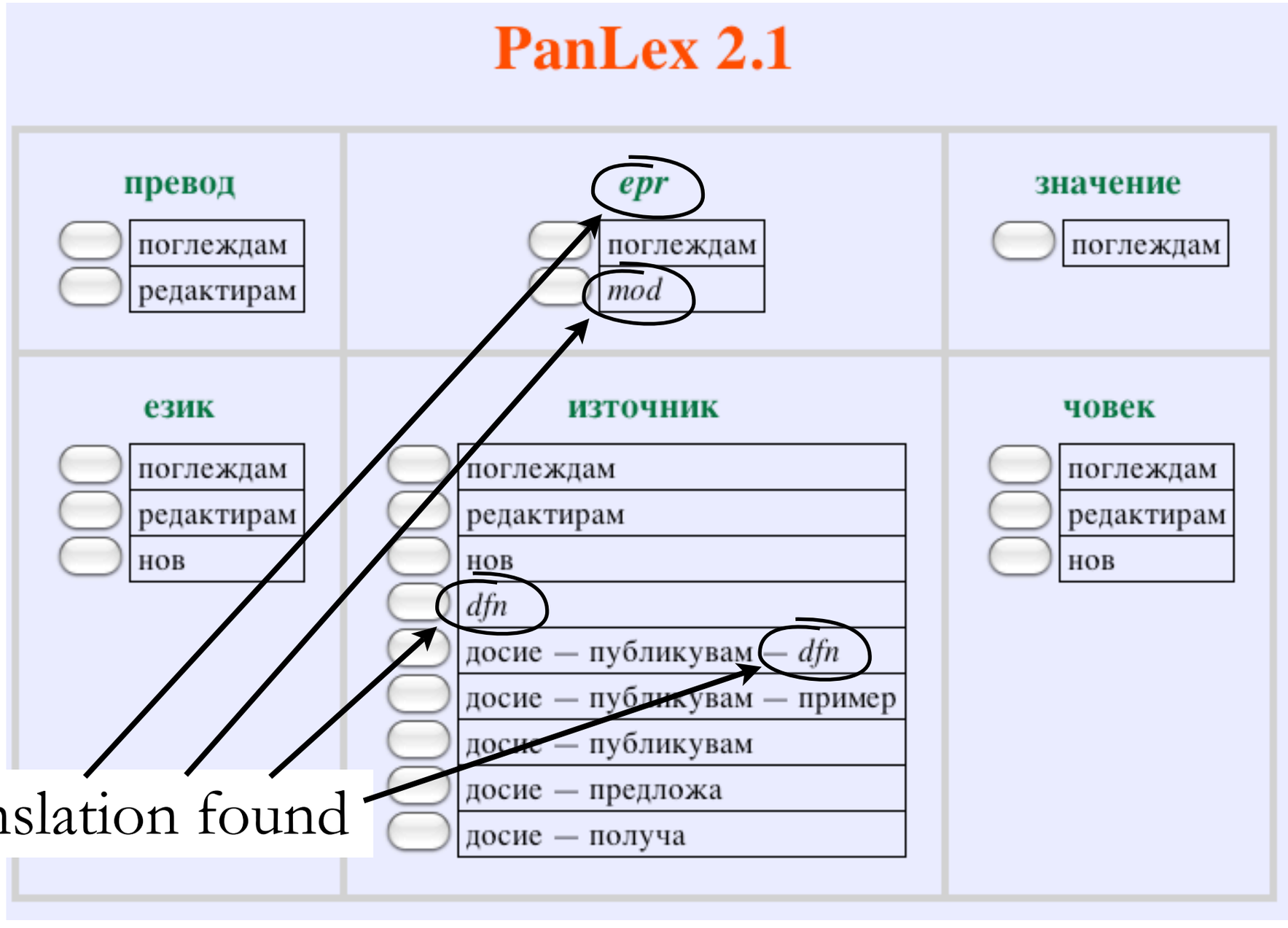
Lemmatic UI



Turkish

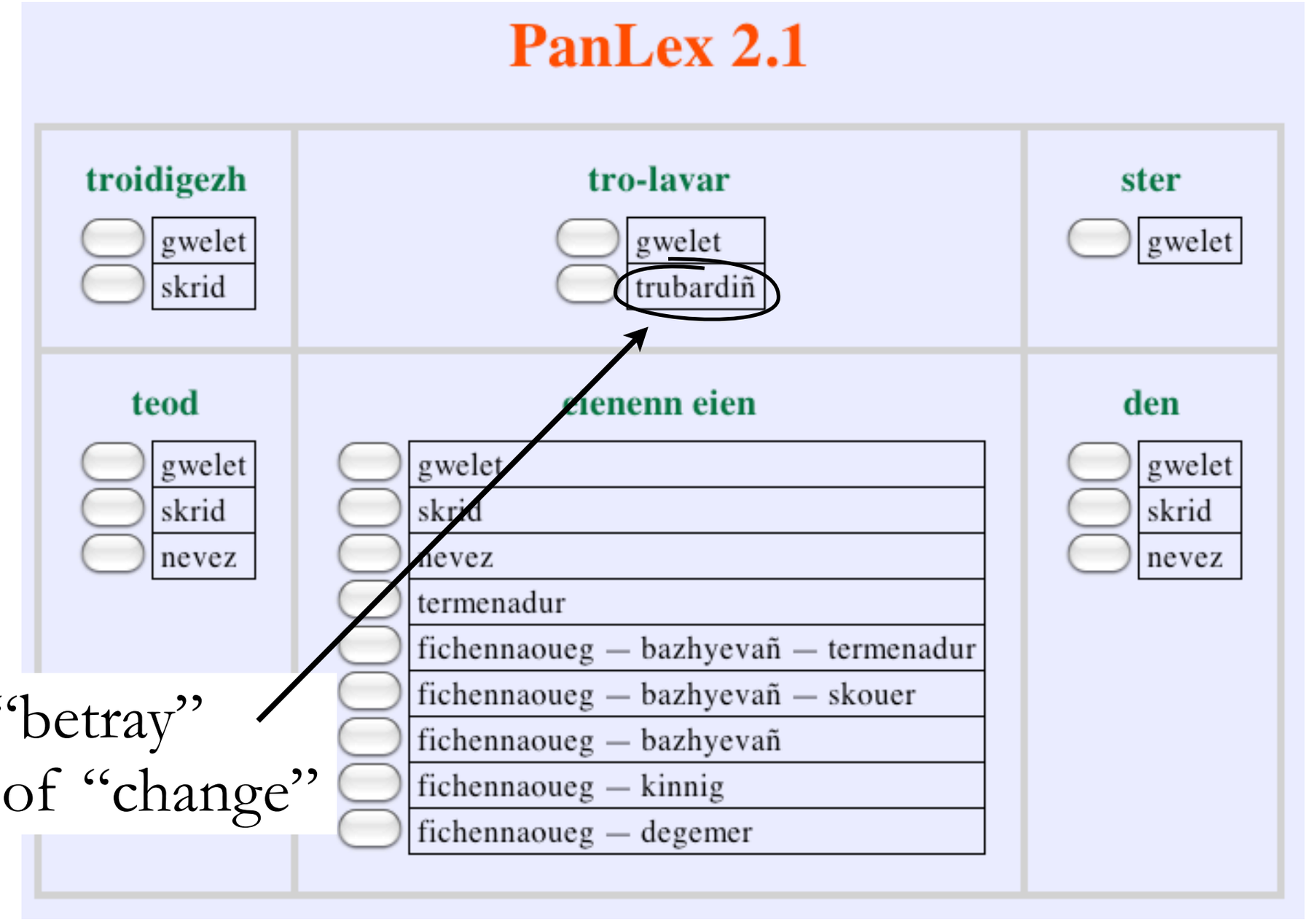
http://panlex.org/u

5. Challenges: Incomplete Data



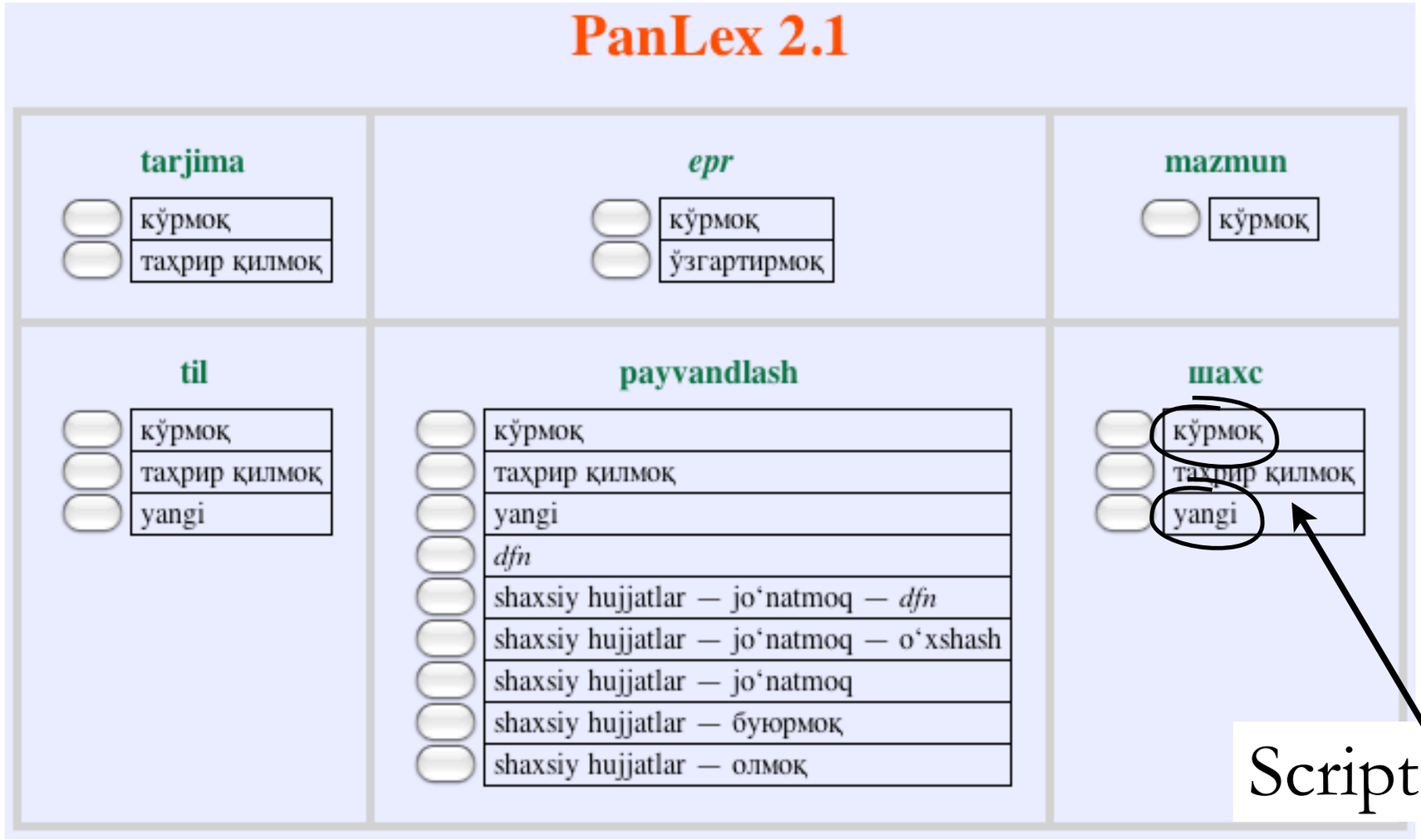
Bulgarian

5. Challenges: Mistranslation



Breton

5. Challenges: Inconsistency



Uzbek

5. Challenges: Intelligibility

PanLex 2.1

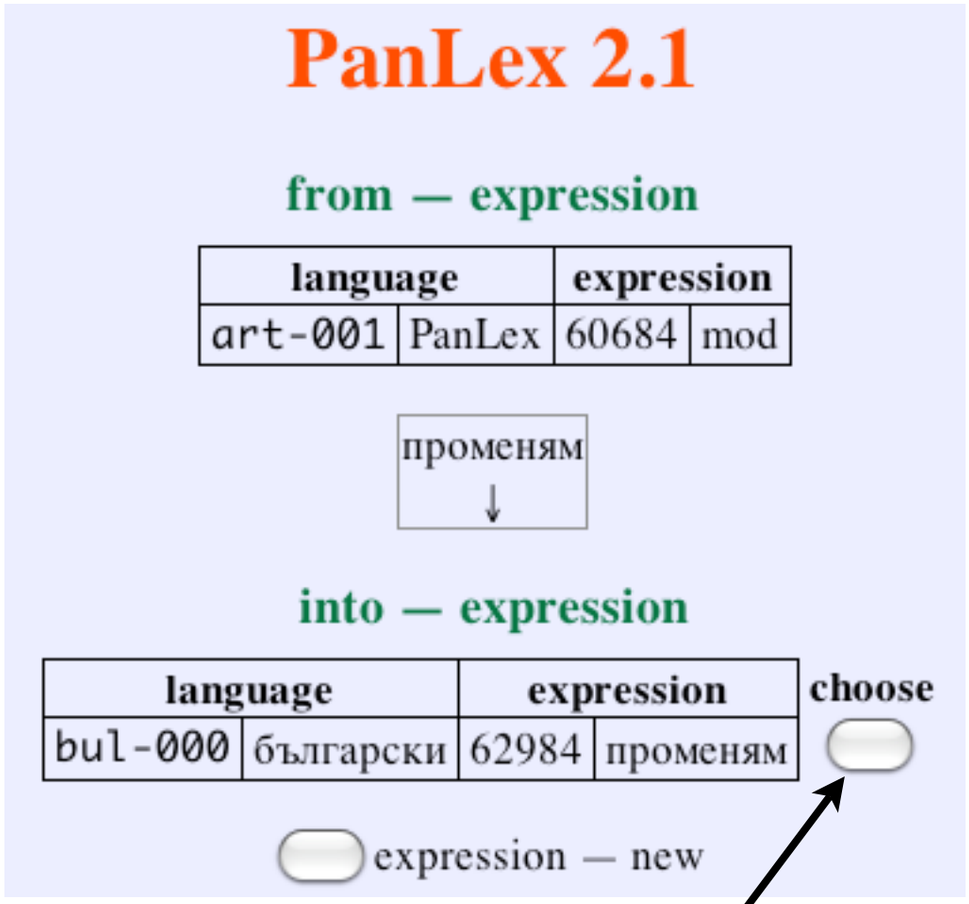
language

number	10
code	ain
variety	0
meaning — expression — many	yes
expression — meaning — many	yes
language — variety — name	アイヌ イタㇰ
source	mul:pool
expression — count	549

expression — text — character — count — see

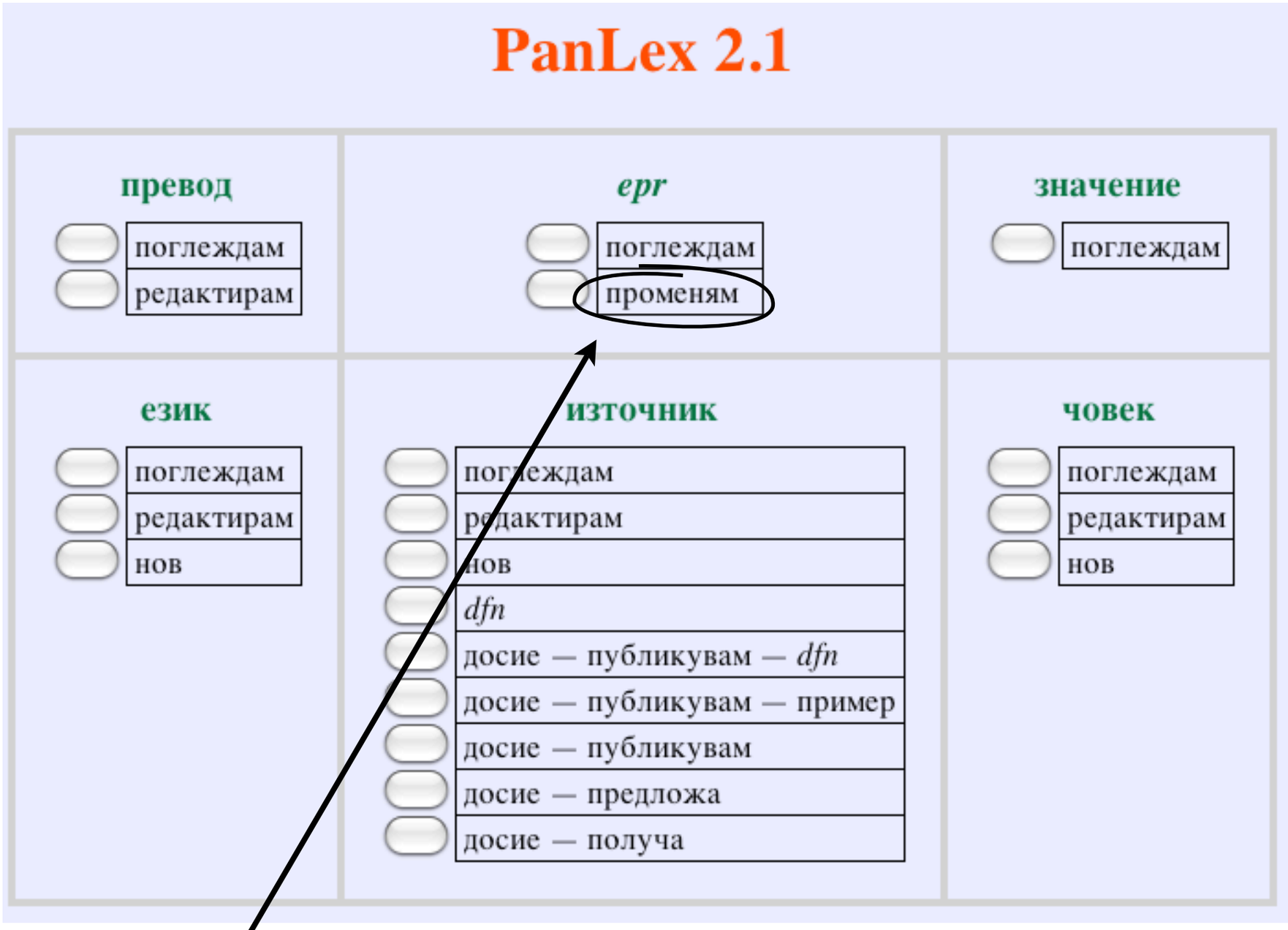
Intent: “Let me see the counts of the characters in the texts of the expressions (in Ainu)”: Clear?

6. Solutions: Patch Holes



Add individual translations

6. Solutions: Patch Holes



New translation used

Utilika Foundation

6. Solutions: Add Massive Content

Indicator	Now	Goal
Resources	600	10000
Language varieties	1300	7000
Expressions	12,000,000	350,000,000
Expression-meaning pairs	27,000,000	1,000,000,000
Translation/synonym pairs	102,000,000	1,000,000,000

Content Sources

Machine-readable lexical resources
Printed lexical resources
Human computation (games, volunteering, piece work)
Structured and unstructured text

6. Solutions: Add Power

To lemmatic communication:
User training
Interfaces to computational grammars and text analyzers
More intelligent inference
Interactive inference

Beyond lemmatic communication:
Morphological data
Syntactic data
Language-independent relational representations

7. Discussion

Is panlingual localization realistic?

Who would be good collaborators?

What applications could benefit?

Principal Collaborators

University of Washington Turing Center:

- Oren Etzioni
- Mausam
- Stephen Soderland
- Marcus Sammer
- Kobi Reiter
- Katherine Everitt

Utilika Foundation:

- Susan Colowick

PostgreSQL Experts, Inc.:

- Quinn Weaver

Utilika Foundation

More Information and Publications

Utilika Foundation:
    http://utilika.org
PanLex project:
    http://panlex.org
PanLex UI prototype:
    http://panlex.org/u
Turing Center:
    http://www.turing.washington.edu

Utilika Foundation