

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



Utilika Foundation

1. Goal: Panlingual Localization

Make any information instantly and freely accessible in any language.

Unicode CLDR: 145

Universal Declaration of Human Rights: 360

Wikipedia: 235

Wikipedia: 235

~~Universal Declaration of Human Rights: 360~~

> Strategy



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



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



tout	heureux	famille	similaire
tout	malheureux	famille	particulier

3. Method: Mine Diverse Lexical Resources

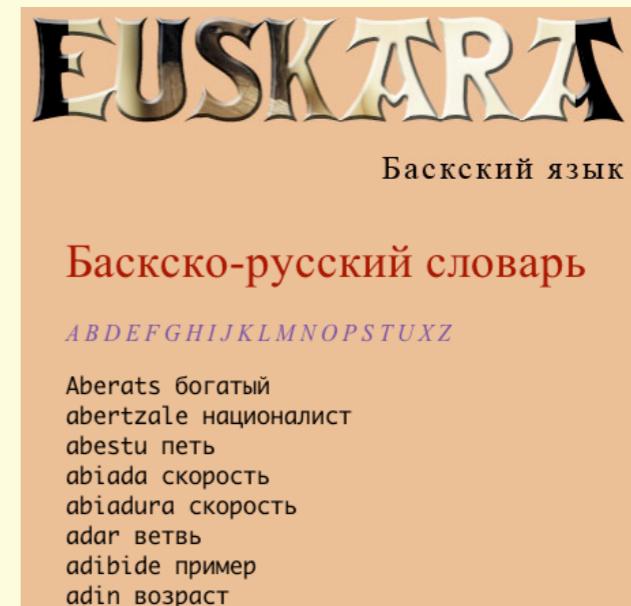


Thesauri

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

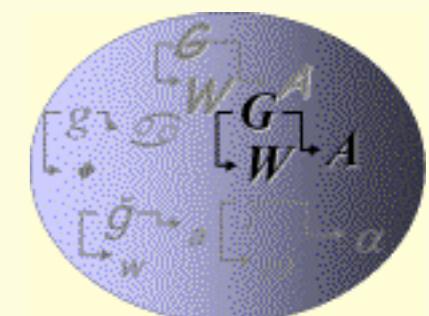
Arrest: ቦንልጋዣ: Tigijauniq: Arrestation

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: ΔΡΩ�በር: Ikitittiniq: Crime d'incendie

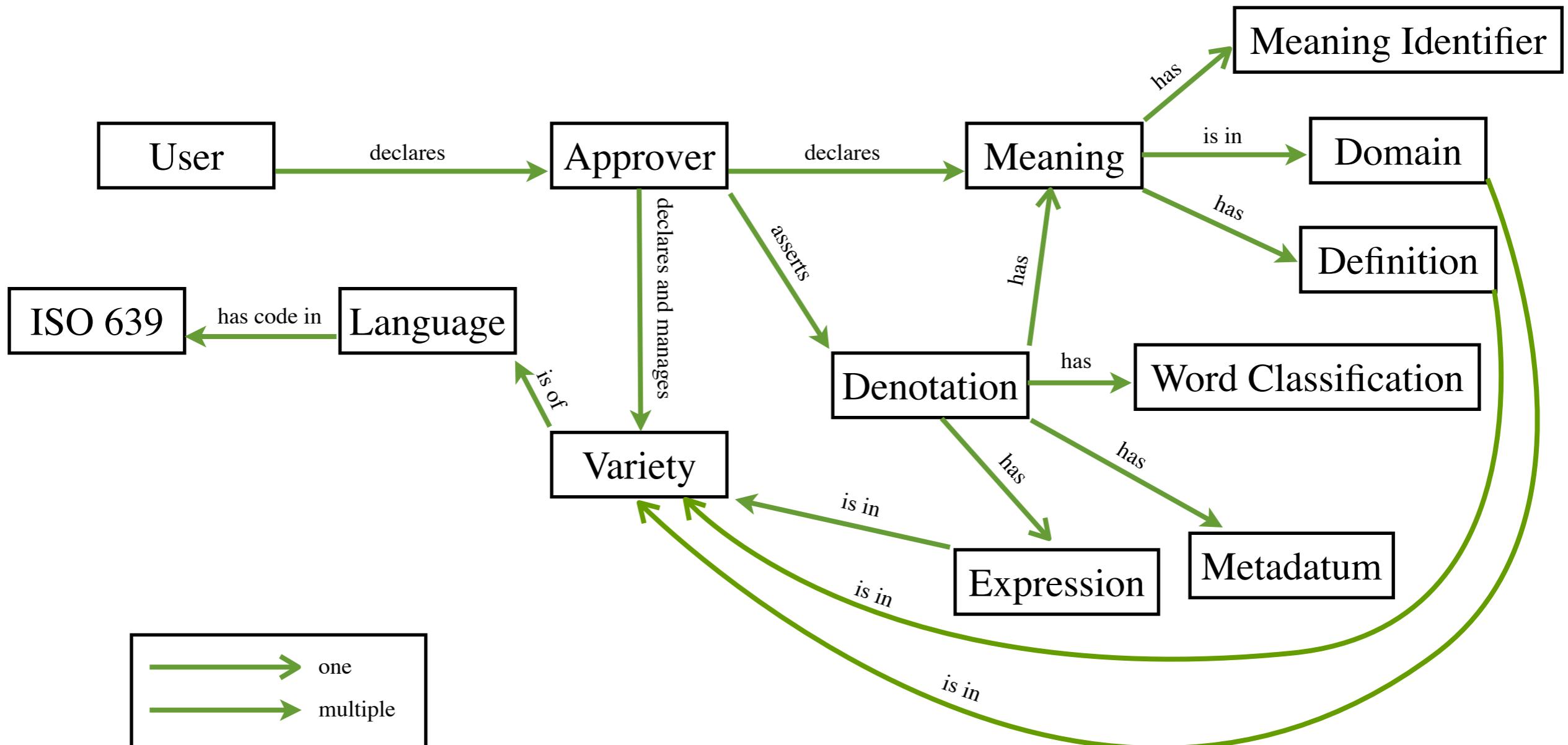
The crime of deliberately setting fire to property. *Criminal Code*, 1996, sections 433-436.

WordNets



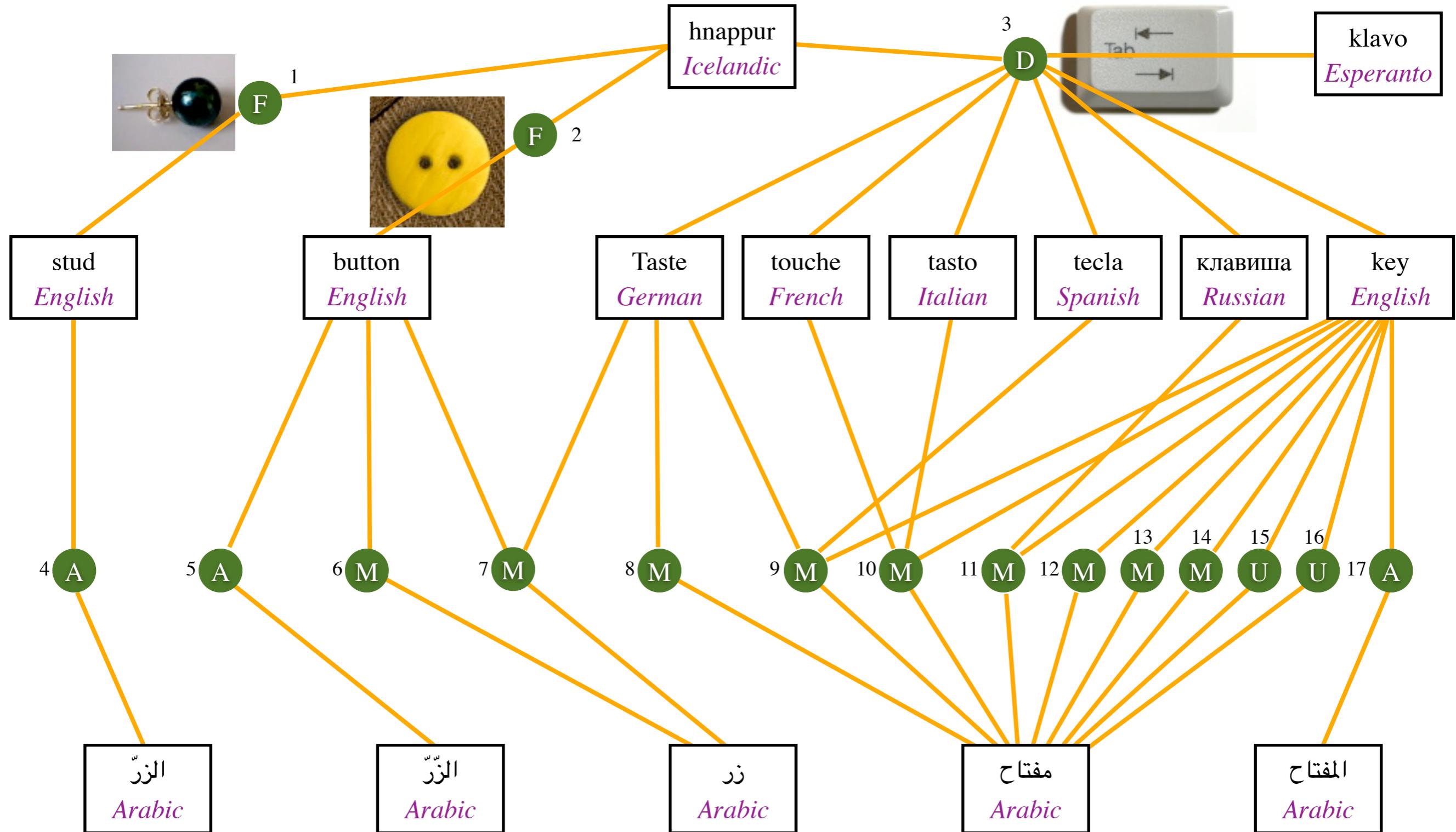
Utilika Foundation

3. Method: Unify Heterogeneous Data



Utilika Foundation

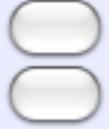
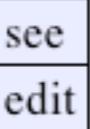
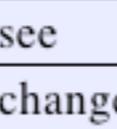
3. Method: Infer New Results



4. Demonstration

Lemmatic UI

PanLex 2.1

translation  	expression  	meaning 
language   	source see edit new definition file — submit — definition file — submit — example file — submit file — propose file — get	person   

English

<http://panlex.org/u>



Utilika Foundation

4. Demonstration

Lemmatic UI

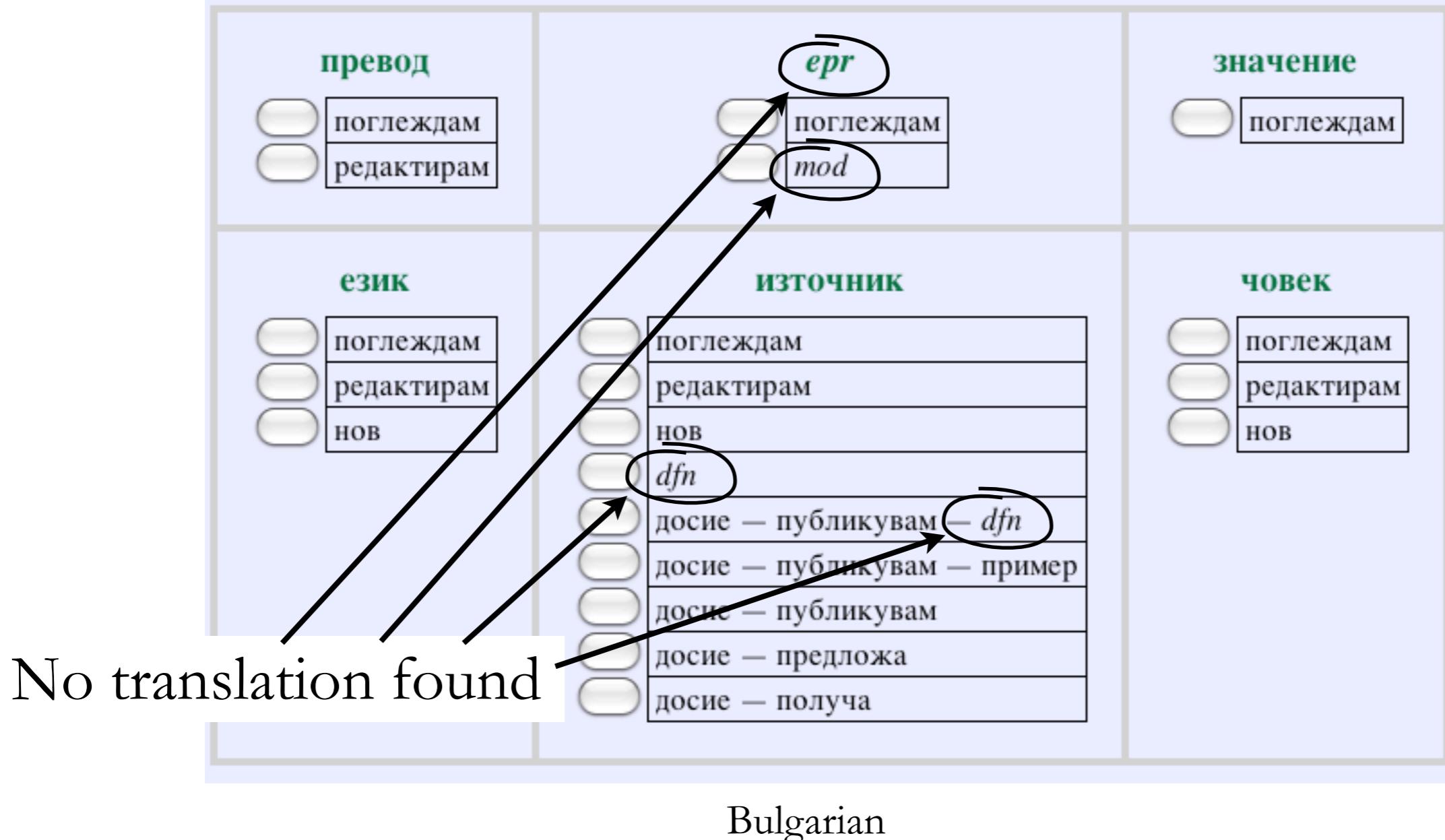
TümSöz 2.1		
çeviri	anlatım	anlam
dil	kaynak	kişi
<p>görmek</p> <p>düzenlemek</p>	<p>görmek</p> <p>değiştirmek</p>	<p>görmek</p>

Turkish



5. Challenges: Incomplete Data

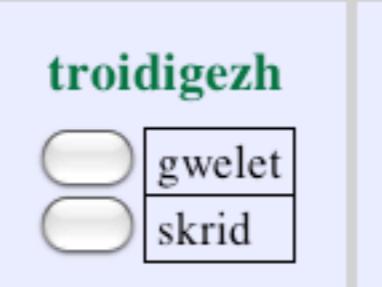
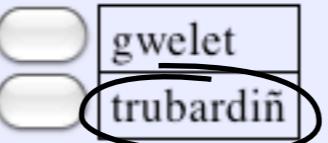
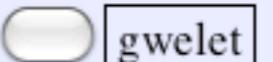
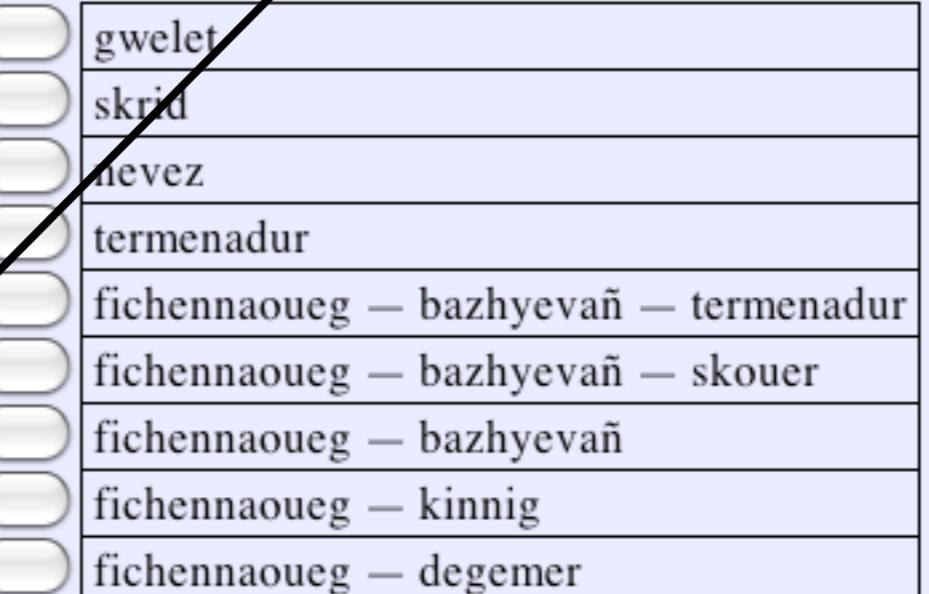
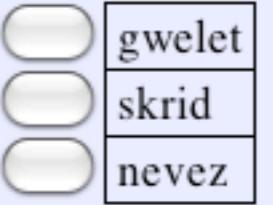
PanLex 2.1



Utilika Foundation

5. Challenges: Mistranslation

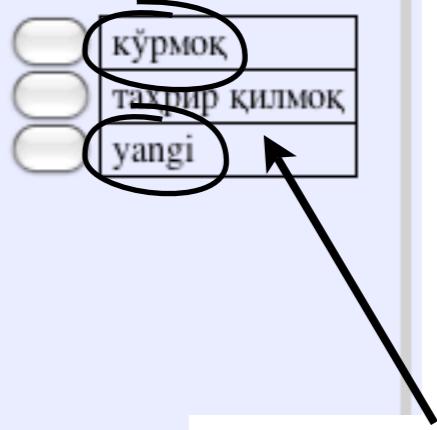
PanLex 2.1

troidigezh  gwelet skrid	tro-lavar  gwelet trubardiñ	ster  gwelet
teod  gwelet skrid nevez	cienenn eien  gwelet skrid nevez termenadur fichennaoueg — bazhyevañ — termenadur fichennaoueg — bazhyevañ — skouer fichennaoueg — bazhyevañ fichennaoueg — kinnig fichennaoueg — degemer	den  gwelet skrid nevez
Means “betray” instead of “change”		Breton



Utilika Foundation

5. Challenges: Inconsistency

PanLex 2.1		
tarjima кўрмоқ тахрир қилмоқ	epr кўрмоқ ўзгартирмоқ	mazmun кўрмоқ
til кўрмоқ тахрир қилмоқ yangi	payvandlash кўрмоқ тахрир қилмоқ yangi <i>dfn</i> shaxsiy hujjatlar — jo'natmoq — <i>dfn</i> shaxsiy hujjatlar — jo'natmoq — o'xshash shaxsiy hujjatlar — jo'natmoq shaxsiy hujjatlar — буюрмоқ shaxsiy hujjatlar — олмоқ	шахс кўрмоқ тахрир қилмоқ yangi
		

Uzbek



Utilika Foundation

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 イタク
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

PanLex 2.1

from — expression

language	expression
art-001	PanLex
60684	mod

променям
↓

into — expression

language	expression	choose
bul-000	български	62984
		променям

expression — new

Add individual
translations



Utilika Foundation

6. Solutions: Patch Holes

PanLex 2.1

превод	epr	значение
<p>поглаждам</p> <p>редактирам</p>	<p>поглаждам</p> <p>променям</p>	<p>поглаждам</p>
език	източник	човек
<p>поглаждам</p> <p>редактирам</p> <p>нов</p>	<p>поглаждам</p> <p>редактирам</p> <p>нов</p> <p><i>dfn</i></p> <p>досие — публикувам — <i>dfn</i></p> <p>досие — публикувам — пример</p> <p>досие — публикувам</p> <p>досие — предложа</p> <p>досие — получа</p>	<p>поглаждам</p> <p>редактирам</p> <p>нов</p>

A black arrow points from the word "нов" in the "език" column to the word "нов" in the "източник" column.

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



Utilika Foundation

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