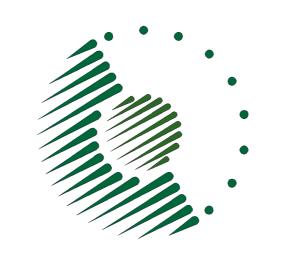
Panlingual Localization

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

עוב 🗸 שֹר מֹים Τια, איייני, אייניק ארמית, Amuzgo de Guerrero, Englisce spræc, Ndéé, Alɔnɑpɑtəwewɑkan, 圆東話, Abáachi mizaa, Aiaha, Gbari, Northwestern Ojibwa, Malayu Ambon, Ndee biyati', Apalaí, אורמית, Apalaí, אורמית, ווייבישה ארמית, אייניק ארמית, אייניק ארמיים, אורמיים, אורמיים aragonesa, Mapudungun, Manado, Cusco Quechua, Hinono'eitiit, Toki Pona, Arawak, مغري ويهم مريي ممريي, Nakoda, অসমীয়া ভাষা, asturianu, Aneityum, asap мацІ, Avestan, অবস্থা, aymar aru, azərbaycanca, Kalenjin, Northwest Gbaya, башкорт теле, بريون ي بادي bamanankan, basa Bali, boarisch, Ďasaá, Baoulé, Bende, Bandjalang, беларуская, बारणा, estremeñu, Betawi, Bena, يريون ي بادي , Classical Nahuatl, भीकी, भोजपुरी भाषा, Bikol, Bislama, Zeêuws, siksiká, Acoli, Bemba, निर्फाल, bosanski, Bondei, विक्थिया मिल्त्री लांगा, brezhoneg, ورودى, वोडो भाषा, Burushaski, Butuanon, буряад хэлэн, 🕫 🚕 🔊 , български, Buli, Batak, Brithenig, Chortí, Garifuna, Caddo, Adangme, català, Ikngerripenhe, Goyogohó:no', binisayang Sinugboanon, Arguni, čeština, Kagayanen, Chamoru, Саtawba, нохчийн мотт, марий, сhahta anumpa, しつかって、GWУ ののようをひれ、Chumash, Мзыкъ словъньскъ, чаващ, Tsetschestaestse, Chikashshanompe, Millingua mon run and Selle murropaparanan numina Koncati Teinlhoot in and navaminan ECE number to the worth Kernewek, Amara, lingua corsa,

Cherepon, "סל∆"ס, Qırımtatar tili, ו Lënape, Deutsch, Tłicho, Dieri, Thu Deyah, Ngas, julakan, Agam, Emb турэн, EDEgbE, Ewondo, Arbore, fi arpitana, Frasche spräke, Frysk, Ba galego, Bench, Babine, گىلكى, chc Ekegusii, Guajá, Gweno, Dinjii Zhu Hmoob, Caribbean Hindustani, Hoj Indonesia, Bima, гІалгІай мотт, Inг.

Make any information instantly and freely accessible in any language.

lyirbal, Allar, Alune, цез мец, l, Ambai, Middle Dutch, Dusun euskara, эвэды торэн, орочон iddle French, Old French, lenga ; na h-Alba, нанай, Gaeilge, Abyá Guaraní, Maléku Jaíka, otsiHerero, Ilonggo, हिंदी, Nesili, , Bo, interlingua, bahasa ıbaylit, kalaallisut, ಕನ್ನಡ,

бежкьа миц, Banjar, کاشر زبان, Bamwe, қазақ, къэбэрдеибзэ, ханты ясан, Kutu, Makonde, Numee, къарай тили, kriolu kabuverdianu, Bontoc, விலியி, Khvarshi, Kickapoo, Gikûyû, тоъфа дыл, kinyaRwanda, кыргыз, Kimbu, Kisi, хакас тили, Guguyimidjir, Klamath-Modoc, Gamilaraay, Konni, zimanê kurdî, Karipúna Creole French, Konkani, перым-коми кыв, कॉकरगी भाषा, коми кыв, kiKongo, 한국어, ньмылан, къарачай-малкъар тил, karjalan kieli, Masiwang, Kuanua, Kölsch, Krisa, Kansa, Karitiâna, Kuanyama, Kuria, къумукъ тил, Kunama, kurdî, Kwakiutl, Karok, Kiput-איספאנייולה, Langi, ਲਹੀਂਦੀ, Lambya, Laka, Lamba, шталала, lengua lumbarda, Bundeli, Musi, Lëtzebuergesch Sprooch, Lakhótiyapi, Lala-Roba, lingaz ladin, lengua lumbarda, Bundeli, Musi, Lëtzebuergesch Sprooch, lüüdikiel, Ganda, Bintulu, Luo, ლაზური ნენა, basa Mathura, Kajin Majel, Blaan, मैथिकी भाषा, മലയാളം, Mandingo, मराठी, ЭІ Maa, Mambai, мокшень кяль, Bilaan, Mekeo, mende, Meru, Motu, Omäeqnomenew, morisyin, Mayo, Middle Irish, Malila, Mambwe-Lungu, Manda, Matengo, Matumbi, Mbunga, myaamia, Mi'gmawi'simg, Brunei Bisaya, Minangkabau, Mískito, Basketo, Amahai, Mwera, Canala, македонски, Malagasy fiteny, Western Maninkakan, Malti, Bata, Bahasa Melayu, Bobot, MeejO, mYtYloN, Atoni, маньси латын, ээээээф, kanien'kéha, limba moldovenească, Bushi, монгод, Mòoré, Mpoto, Majang, Yosondúa Mixtec, Misima-Paneati, Maricopa, reo Māori, Bubi, кырык мары йылмы, Maranao, bahasa Melayu, Bukat, Bwaidoka, mvskokē, Murrinh-Patha, Ihéngua mirandesa, Kala Lagaw Ya, मारवाङ्ग, क्रिक्ट, Bangala, Буряад хэлэн, с'ор, Рима, Mundurukú, эрзянь кель, Masaba, nawatlahtolli, Bargu Buriat, 台灣話, Bilin, Babuza, Indonesia Buli, lengua nnapulitana, Carolinian, Chiga, Khoekhoegowab, dorerin Naoero, Dine bixasi, isiNdebele, Ndali, Ndarab, Ndonga, Plattdüütsche Sprook, Chuukese, Chippewa, Kavalan, Cuyonon, Nemi, नेपाली, Taita, नेपाल भाषा, Nez Perce, Ngoni, Nyiha, ko e vagahau Niue, Dagaari Dioula, нивхгу диф, Nederlands अक्षेत्र. Zazaki. Gela कि अपिता Damar, nynorsk, Ngindo, ногай тили, , Omaha-Ponca, South Efate,

norskr, Dobu, Aputai, Novial, langue nouormande, c\b\h, sePedi, West Damar, chiCheŵa, Rukai, Nyamwezi, Nyunga, Nyakyusa-Ngonde, lenga d'òc, E Tohono O'odham, Northwest Alaska Inupiatun, Estonian Sign Language, ଓଡ଼ିଆ ଭାୟା, afaan Oromoo, ирон æвзаг, Old Saxon, لمران عشماني , Old Turkish, tekoi ra Belau, Pawnee, Pangwa, langwa picarde, Pennsilfaani-Deitsch Schpropeli, Plautdietsche Sproak, المارس في Eastern Frisian, Norfuk, Powhatan, Pijin Futuna-Aniwa, Gayo, polski, Pohnpeian, podruguês, Guianese Creole Frencio, Sayula Popoluca, Neshnabémwen, Malecite-Passamaquoddy, Prakrit, Parsi-Dari, Pra

ang, ਪੰਜਾਬੀ, Papiamentu, Siraya, As, पालि भाषा, lenga piemontèisa, , Gedaged, Gebe, Geser-Gorom, Paiwan,

română, Kriol, руськи язиг Sahtúot'ine Yati', cmiique. slovenščina, åarjelsaemien;

Runasimi, Quileute, Quenya Panansi Macri Kalo Sirani Romans Kon Sinto Caló Kulderašicko Romani čhib, Patpatar, Rungwa, Ata Manobo, Hinukh Cano-Gofa-Dawro, Gorontalo, Losua rumantscha, Ghari, Romani čhib, , limba meglenoromană, русский, yângâ tî săngō, Gane, сक काम काम, संस्कृत भाषा, Ha, संधाकी, Safwa, Sangu, lingua siciliana, Scots leid,) 'afó, Mende, Ёохів, ca'мь кйлл, Sindarin, Southern Sierra Miwok, Seraiki, Ancient Hebrew, Pamplona Atta, ślonsko godka, slovenčina, škielâ, gagana fa'a Samoa, sää mkiöll, chiShona, درنوی, Soninkanxaane, Sihan, Soomaaliga, Anuta, seSotho, Hanunoo, español, shqipe,

amáka, Sranantongo, Campidanese Sardinian, српски, Waray Sorsogon, siSwati, Fräiske Sproake, x^wsenəčqən, Suba, Shubi, Sukuma, Surigaonon, basa Sunda, Sumbwa, Sumerian, ლუმნუ ნინ, Gullah, طَيْدُرُ, svenska, Sap Maco Del Mar Huave, Kiswans 👡 കം, табасаран чІал, гео Tahiti, ფამს , Hunzib, Central Tarahumara, tatar tele, Torres Strait Creole, San Luís Potosí Huastec, ভোষ্টা, **თулобвајол**, Зеох., Teso, lia-tetun, Tewa, точики даlog, ภาษาไทย, Tharaka, ቲ %, ት %. 💘 thIngan Hol, Lingít, тольшэ зывон, عراضي, Taino, chiTonga, faka-Tonga, Tok Pisin, Tupinambá, Tonkawa, Даков, Тагоко, Tausug, Tsimshian, seTswana, Sabu, xiTsonga, ៉្នែ, ភាមាគឺការ, тати, türkmen, Türkçe, Rurutuan, 'gana Tuvalu, Igala, Twi, Tokharian B, Thracian, тыва дыл, Chamula Tzotzil, tamaziyt, Western Tzutujil, Interglossa, Ili'uun, удин муз, удмурт кыл, Ugaritic, ويَدِي فِي فِي فِي وَهِي بِهِر ykpaïhcska, Meriam, اردي , ykpaïhcska, Meriam, اردي , ykpaïhcska, Meriam, اردي , ykpaïhcska, Meriam, اردي , ykpaïhcska, Meriam, الوعي إلى إلى الو-Southern Paiute, oʻzbek, léngua vèneta, Talur, tshi Venda, vepsän kel', tiếng Việt, Vlaams, Martuyhunira, Mbabaram, Minigir, Volapūk, võro kiil', vad'd'a tšeeli, Vunjo, Wappo, Wanda, Wanji, x'ik zik, Warlpiri, Serbšćina, Imroing, Irarutu, Wik-Ngathana, Wik-Mungkan, Western Jacalteco, Middle Welsh, Lingaedje walon, Jawe, kallaama wolof, Wiradhuri, Bungu, 吳語/吳语, Wyandot, Wangaaybuwan-Ngiyambaa, хальыг келн, Xavánte, Kamba, Middle Breton, Dacian, isiXhosa, Yamdena, Lycian, Lydian, дыбадост 160 бобы, Рокын Ж. Sudovian, Katcha, Kadugli-Miri, Tokharian A, Umbrian, Kamara, North

Babar, Yao, יידיש, Yindjibarndi, Yokuts, èdè Yorùbá, ненэця' вада, ñc'engatú, Maaya T'aan, Yuchi, Kerewe, 廣東 коді, козгаева, крепе, каріпдатнагаў. Кіпагаў-А, Кurukh, Shambala, Kedang, 'Auhelawa, Kusaal, Kove, Kerinc Lenje, Leipon, Lahanan, Likan Li'o, Lampung, Lau, Lamogai, Hano Lamboya, Lou, Kréyol La Lwizyàn, Indonesii Universal Declaration Manobo, Maisin, Dizi, Mengen, Mela, Mambae, Mailu, Indonesia Baru, Mor Islands Mor, To'abaita, Emae, Muna, Manobo, Maisin, Dizi, Mengen, Mela, Mambae, Mailu, Indonesia Baru, Mor Islands Mor, To'abaita, Emae, Muna, Manobo, Maisin, Dizi, Mengen, Mela, Mambae, Mailu, Indonesia Baru, Mor Islands Mor, To'abaita, Emae, Muna, Manobo, Maisin, Dizi, Mengen, Mela, Mambae, Mailu, Indonesia Baru, Mor Islands Mor, To'abaita, Emae, Muna, Manobo, Maisin, Dizi, Mengen, Mela, Mangareva, Cotabato Manobo, Solomon Islands Mono, Mola, Manam, Mamboru, Mouk-Aria, Nyamwanga, Modan,

Ndengereko, Nengone, Ng Pukapuka, Paulohi, Plateau Tanzania Kara, Rejang, Re

of Human Rights: 360 ya, Nukuoro, Nama, bokmål, Nyambo, :a, Tuamotuan, Penrhyn, occitan ancial.

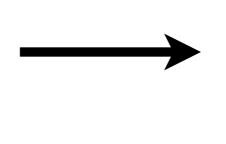
, Zinza, Kaurna, isiZulu, ?, Kisar, Koongo, upulau, Wampar, Lavatbura-Lamusong, Central Mazahua, Western Bukidnon quesan, Manggarai, North Marquesan, currunggurr, South Ndebele, Nauna, Iochi, Ottawa, Oyda, Southern Pashto, "ichua, Logooli, Réunion Creole French,

stuman, Roviana, Central Okinawan, Sandawe, Sasai Sassarese Sardinian, Gallurese Sardinian, Kayraboro Senni Songhai, Sebat Bet Gurage, Shi, , 1907, Schmat, So'a, Sengseng, Tarpia, Sumo-Mayangna, Susu, Suaroa, دري مه المحرية, Yami, Atayal, Mandara, Tboli, Tetun Dili, Tigak, Timugon Murut, Teop, Numbami, Sika, Sikaiana, Sekar, Sera Tunjung, Tokelau, Tikopia, Teany Suchek, Lenakel, Tontemboss, Tobian, Tupí, Tsou, Tsishingini, Tooro, Tumbuka, Vaghua, Southeast Ambrym, Tela-Masbuar, Termanu, Tugun, Ubir, Ujir, Umbundu, Urubú-Kaapor, Kulon-Pazeh, Southeast Babar Veluws, Vinza, East Masela, Mainfrank, Sh. Varisi, Wandamen, Watubela, Winaray, Ho-Chunk, Muduapa, Warji, Wolio, Wallisian, Wanukaka, Wogeo, Waropen, Wuvulu-Aua, Aquitanian, Kambera, Mori Bawah, Kanakanabu, Soga, Puyo-Paekche, Kasem, Saisiyat, Yámana, Yapese, Yakamul, Zayse-Zergulla, Ngazidja Comorian, Magori, Zigula, Shiwi'ma, Qaqet, Belait, Dayak Bakatiq-Sara/Riok, Bugotu, Banggai, Popalia, Barok, Binukid, Balangaw, Bonerate, Kei Besar Elat, Bunun, Bilibil, Bilur, Basai, Chru, Phan Rang Cham, Ida'an, Fataluku, Diodio, Dawera-Daweloor, Casiguran Dumagat, Dai, Doura, Bundu Dusun, Emplawas, Ende, Enggano, Sye, Gaddang, Kubokota, Gumawana, Haku, Hainan Cham, Iaai, Iban, Inibaloi, Amganad Ifugao, Batad Ifugao, Keley-I Kallahan, Ilongot, Iranun, Dibagat-Kabugao-Isneg, Binongan Itneg, Yabem, Kilokaka, Japhug Rgyalrong, Kayapa Proper Kallahan, Gabadi, Kemak, Komering, Kia, Kokota, Guinaang Kalinga, Kalagan, Kandas, Kuni, Kelabit, Kaidipang, Tungag, Kara, Luang, Luqa, Lihir, Nguna, Merei, Lamalera, Loniu, Naman, Dibabawon Manobo, Ilianen Manobo, Sarangani Manobo, Tigwa Manobo, Melanau, Maleu, PanLex, Ma'anyan, Matukar, Mokilese, Makasae, Mwotlap, Mamanwa, Madak, Molima, Mortlockese, Cheke Holo, Mansaka, Miyako, Marovo, Peterara, Moken, White Hmong, Nakanai, Nalik, Nêlêmwa, Aiwoo, Nanggu, Nila, Lala, Nage, Luangiua, Kadorih, Pingilapese, Vaeakau-Taumako, Southwest Palawano, Paamese, Kaulong, Puluwatese, Ngaibor, Proto-Cen E Malayo-Polynesian, Proto-Cen Malayo-Polynesian, Proto-Mulayo-Polynesian, Proto-Oceanic, Proto-Austronesian, Proto-Southern Mindanaon, Pondok, Baliledo, Old Chinese, Proto-Micronesian, Proto-Chuukic, Riwo, Nese, Tape, Namakir, Orkon, Näti, Austronesian?, Austronesian?, Kayan, Roma, Roro, Amami, Simbo, Saliba, Botolan Sambal, Punan Kelai, Mapun, Siar, Sakao, Lamaholot Ile Mandiri, Samal, Vilirupu, Singhi, Saragani Islands Sangil, Nahavaq, Solos, Taiof, Serua, Natqgu, Santa Ana, Satawalese, Siocon Subanon, Serili, Savosavo, Sangir, Sindangan Subanun, Megiar, Tanimbili, Tiang, Kalamian Tagbanwa, Tanga, Tagbanwa, Tolo, Sou Amana Teru, Soboyo, Avava, Kwamera, Tanema, Torau, Sivisa Titan, Teun, Tonsea, Uruava, Nembao, Ura, West Uvea, Vano, Neve'ei, Wedau, Perai, Wejewa, Woleaian, Kenyah, Biga, Northern Kankanay, Buyang, Yakan, Melawan, Melayu, 350. 🗓 , lingua osca, Galatian, fulginate, santostefanese, ternano, caccianese, patermese, patermitano, sanremasco, bregagliotto, milanese, trentino, holupaka, français acadien, Duh-lian, Žemaitiu, võro kiil, Tarandine, Basa Banyumasan, 古文, тарашкевіца, Nedersaksisch, Lucumí, Mundani



2. Strategy: Lemmatic Communication

Все счастливые семьи похожи друг на друга, каждая несчастливая семья несчастлива посвоему.



все счастливый семья похожий все несчастливый семья особенный



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



toutheureuxfamillesimilairetoutmalheureuxfamilleparticulier



3. Method: Mine Diverse Lexical Resources

Thesauri

Wiktionaries



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

NI: vaktaal Sv: fackspråk BT γλώσσες

Dictionaries



Баскский язык

Баскско-русский словарь

ABDEFGHIJKLMNOPSTUXZ

Aberats богатый abertzale националист abestu петь abiada скорость abiadura скорость adar ветвь adibide пример adin возраст

Locale Repositories

SE (English: Sweden)

An tSualainn
isveç
isveç
isveç
Iswidhan
Rootsi
Ruotsi
Ruotta

de ·

gsw '

Schweden

Schweede

Glossaries

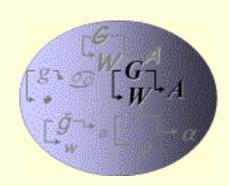
Arrest: ∩リケトσ ና tigujauniq: Arrestation

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: Δρ∩^c∩σ^{co}: Ikitittiniq: Crime d'incendie

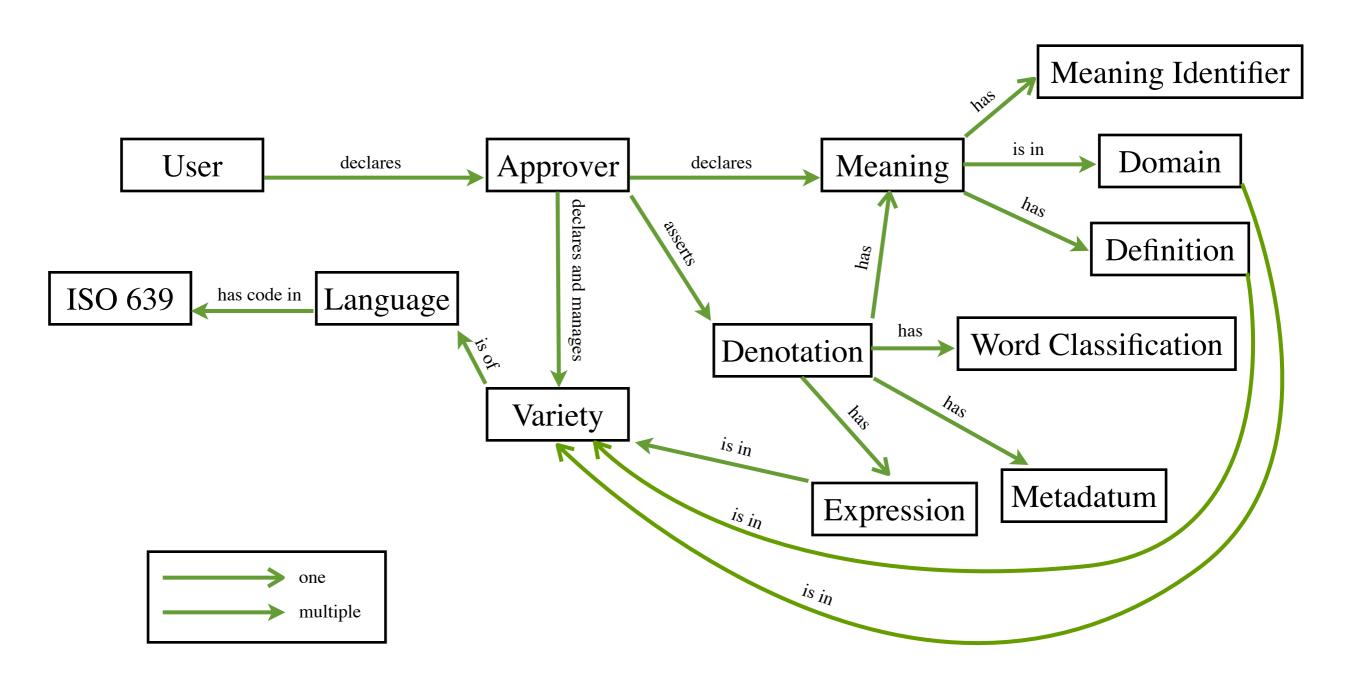
The crime of deliberately setting fire to property. *Criminal Code*, 1996, sections 433-436.

WordNets



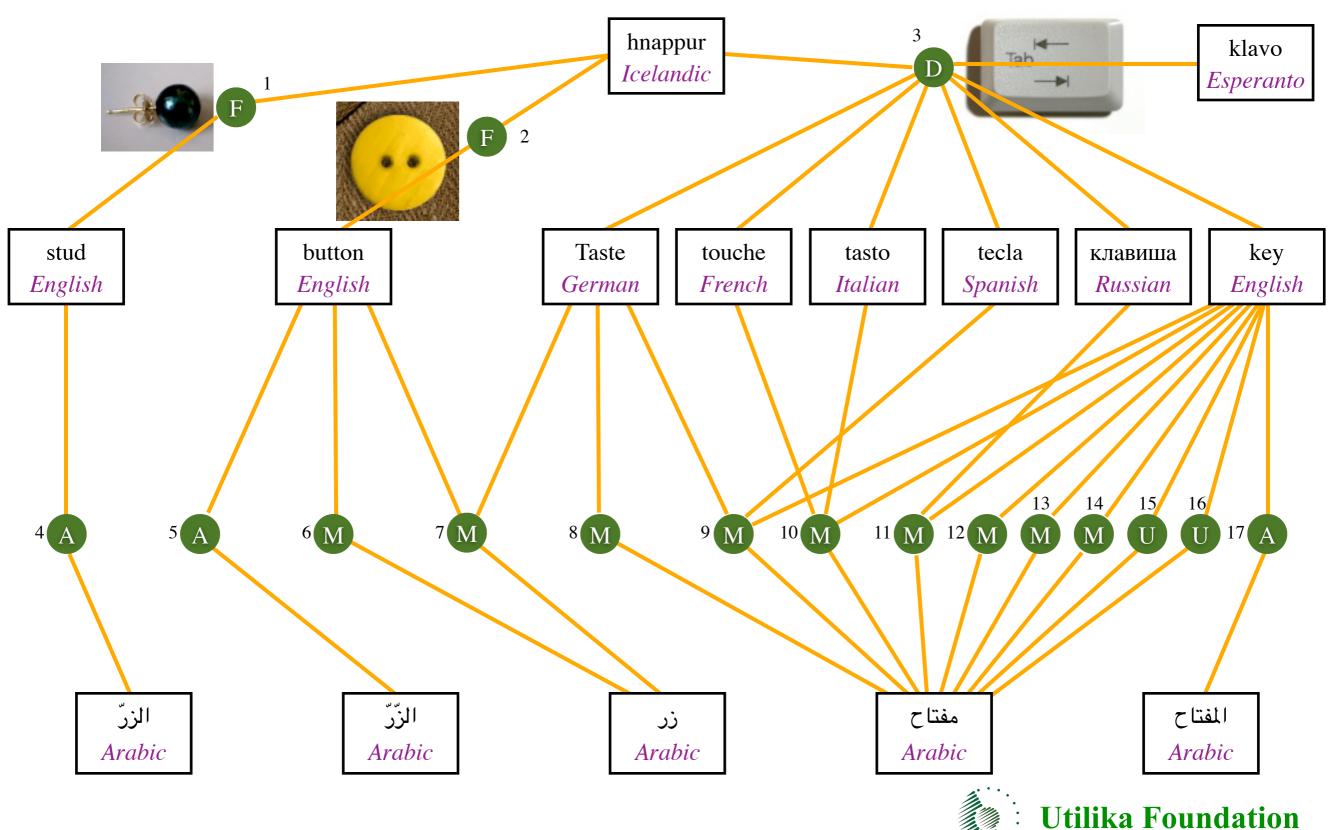


3. Method: Unify Heterogeneous Data





3. Method: Infer New Results



4. Demonstration

Lemmatic UI

PanLex 2.1				
translation	expression see change	meaning		
see edit new	see edit new definition file — submit — definition file — submit — example file — submit file — propose file — get	person see edit new		

English

http://panlex.org/u



4. Demonstration

Lemmatic UI

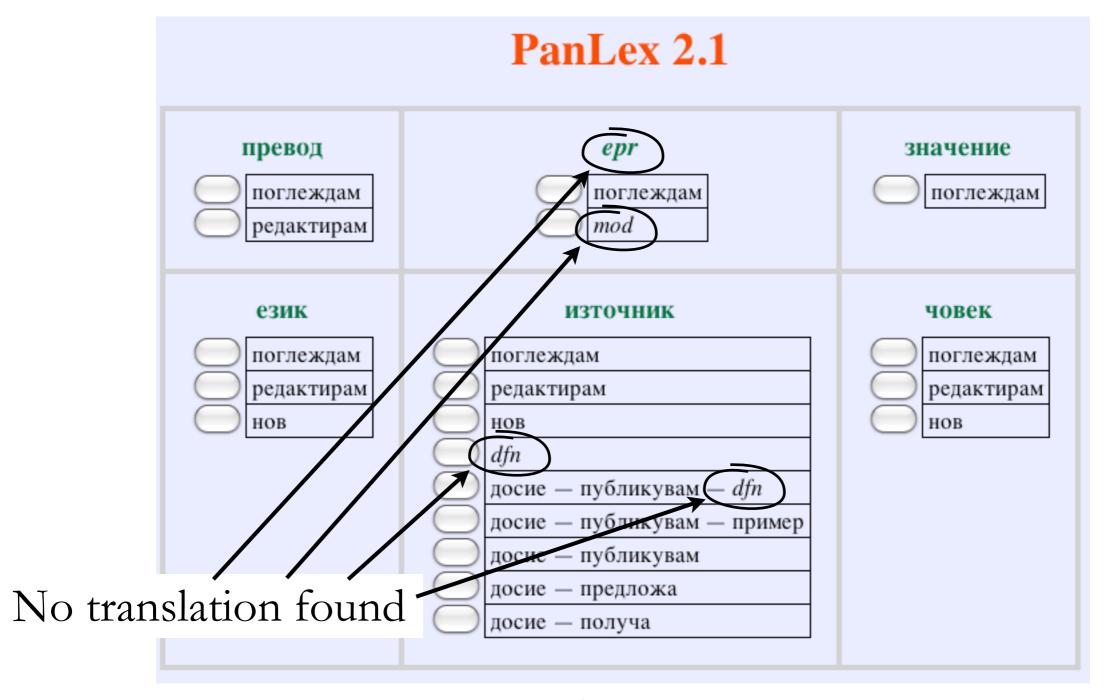
TümSöz 2.1					
çeviri görmek düzenlemek	anlatım görmek değiştirmek	anlam görmek			
dil görmek düzenlemek yeni	kaynak görmek düzenlemek yeni tanım dosya — göndermek — tanım dosya — göndermek — örnek dosya — göndermek dosya — göndermek dosya — önermek dosya — almak	kişi görmek düzenlemek yeni			

Turkish

http://panlex.org/u



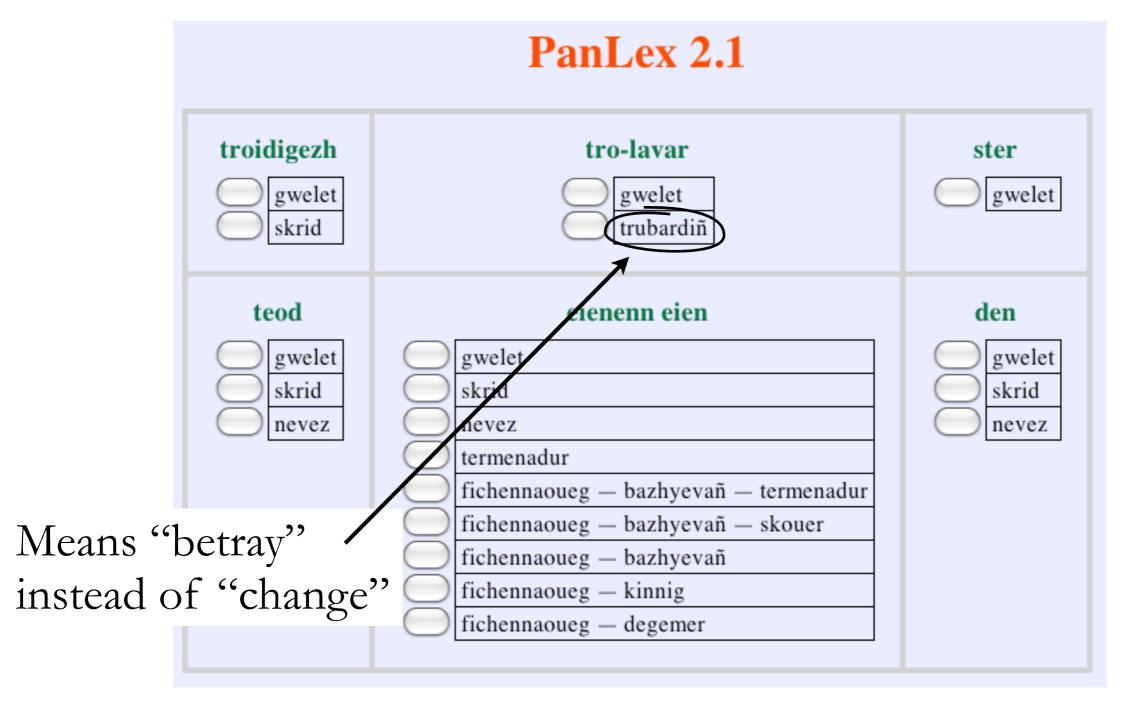
5. Challenges: Incomplete Data



Bulgarian



5. Challenges: Mistranslation



Breton



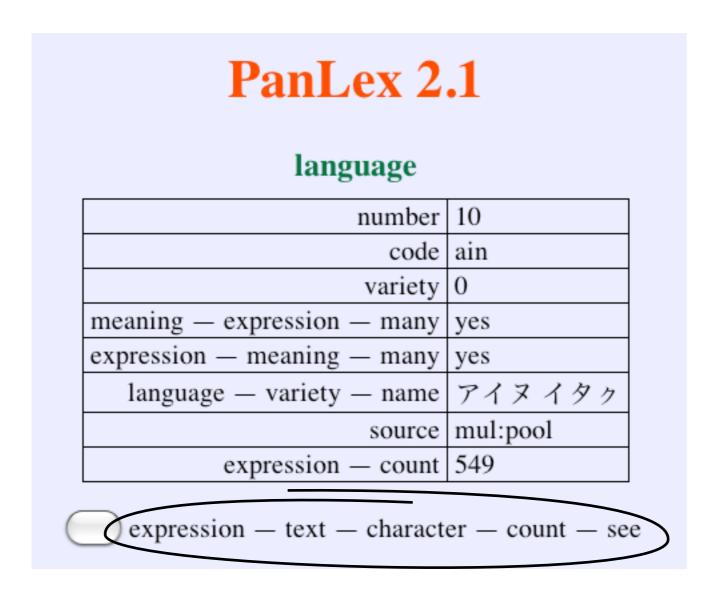
5. Challenges: Inconsistency

PanLex 2.1				
tarjima	<i>ерг</i>	mazmun Кўрмоқ		
til кўрмоқ тахрир қилмоқ yangi	рауvandlash Кўрмоқ тахрир қилмоқ yangi dfn shaxsiy hujjatlar — joʻnatmoq — dfn shaxsiy hujjatlar — joʻnatmoq — oʻxshash shaxsiy hujjatlar — joʻnatmoq shaxsiy hujjatlar — буюрмоқ shaxsiy hujjatlar — олмоқ	жўрмок тахрир қилмок yangi		

Uzbek



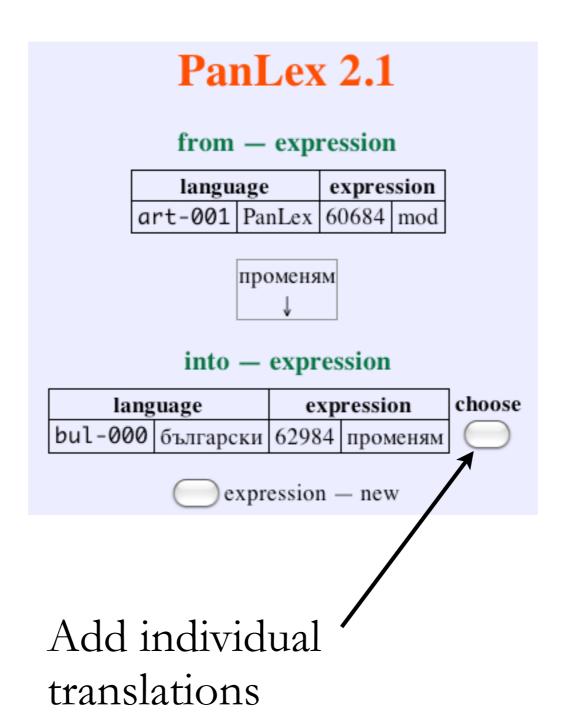
5. Challenges: Intelligibility



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

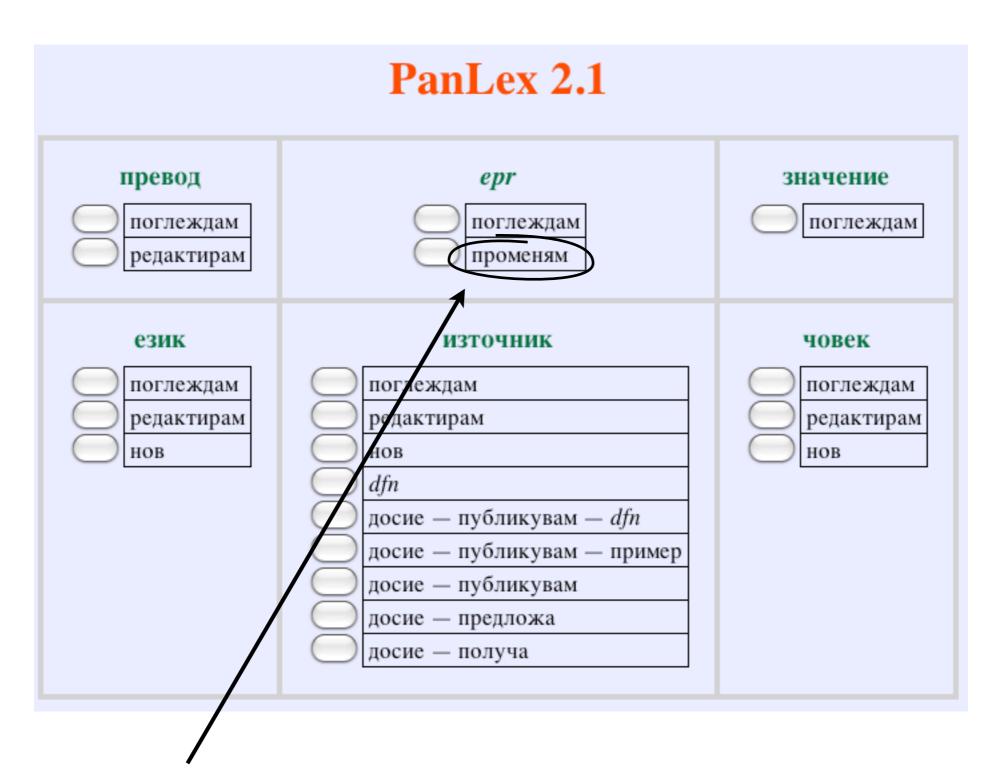


6. Solutions: Patch Holes





6. Solutions: Patch Holes



New translation used



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



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
```

