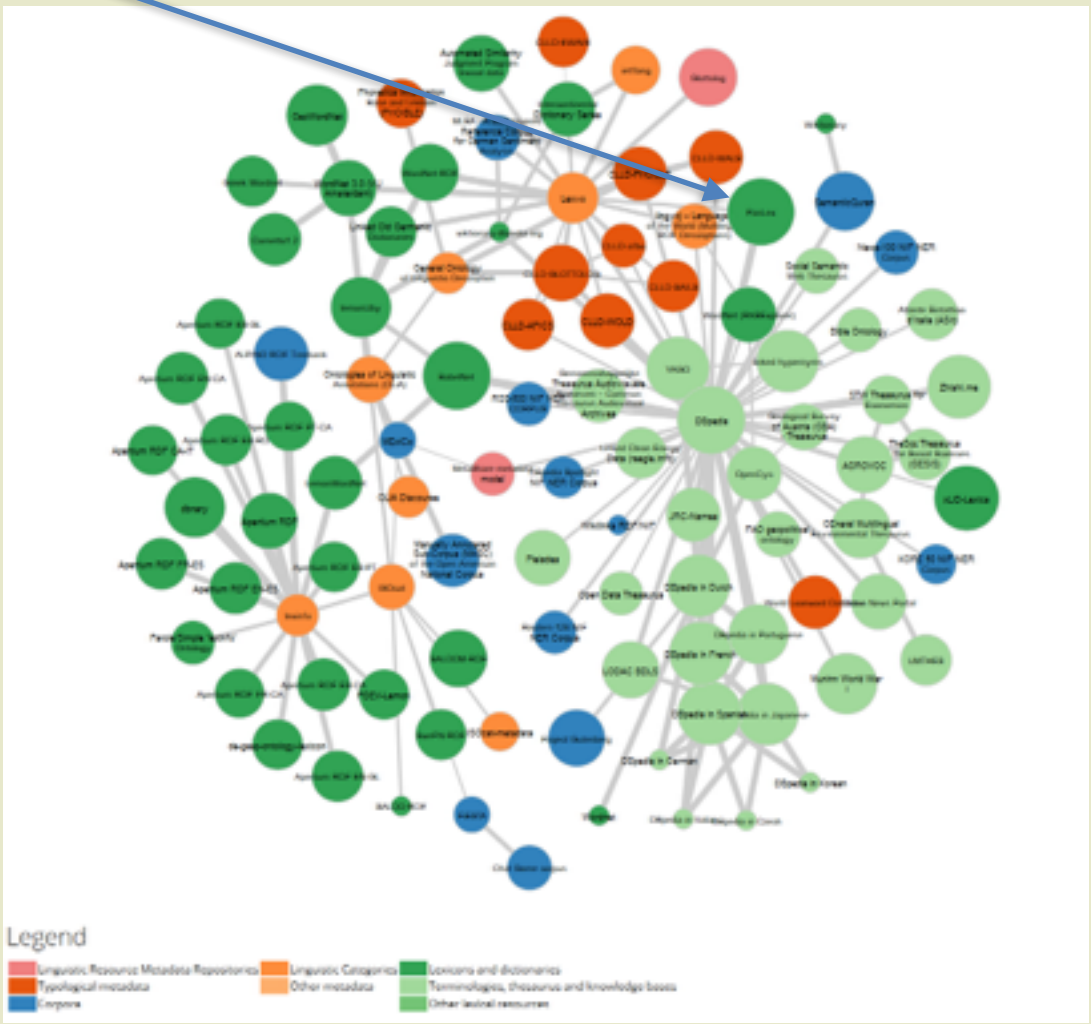
PanLex: Building a Resource for Panlingual Lexical Translation

David Kamholz, Jonathan Pool, and Susan Colowick

Background

- PanLex project started in 2005 at the University of Washington Turing Center
- Since 2012, sponsored by The Long Now Foundation in San Francisco (longnow.org)

PanLex in the LLOD cloud



Introduction

- Goal: Enable panlingual lexical translation: the translation of lexemes among *all* human languages in the world
- Hypothesis: Lexemic translation data provide greatest return on investment in achieving panlingual translation
  - widely available for thousands of languages, as opposed to rich lexical data and corpus data
  - lemmatic communication has various use cases

Introduction

- Project focus:
  - procuring content
  - conversion of data into consistent structure
  - maintaining high-quality dataset
  - making data publicly available
- Left to others:
  - application development
  - translation inference (inferring unattested translations from the existing dataset)

Coverage

- As of March 2014, PanLex database contains:
  - 20 million lexemes (“expressions”) in about 9000 language varieties
  - 1.1 billion pairwise translations among lexemes

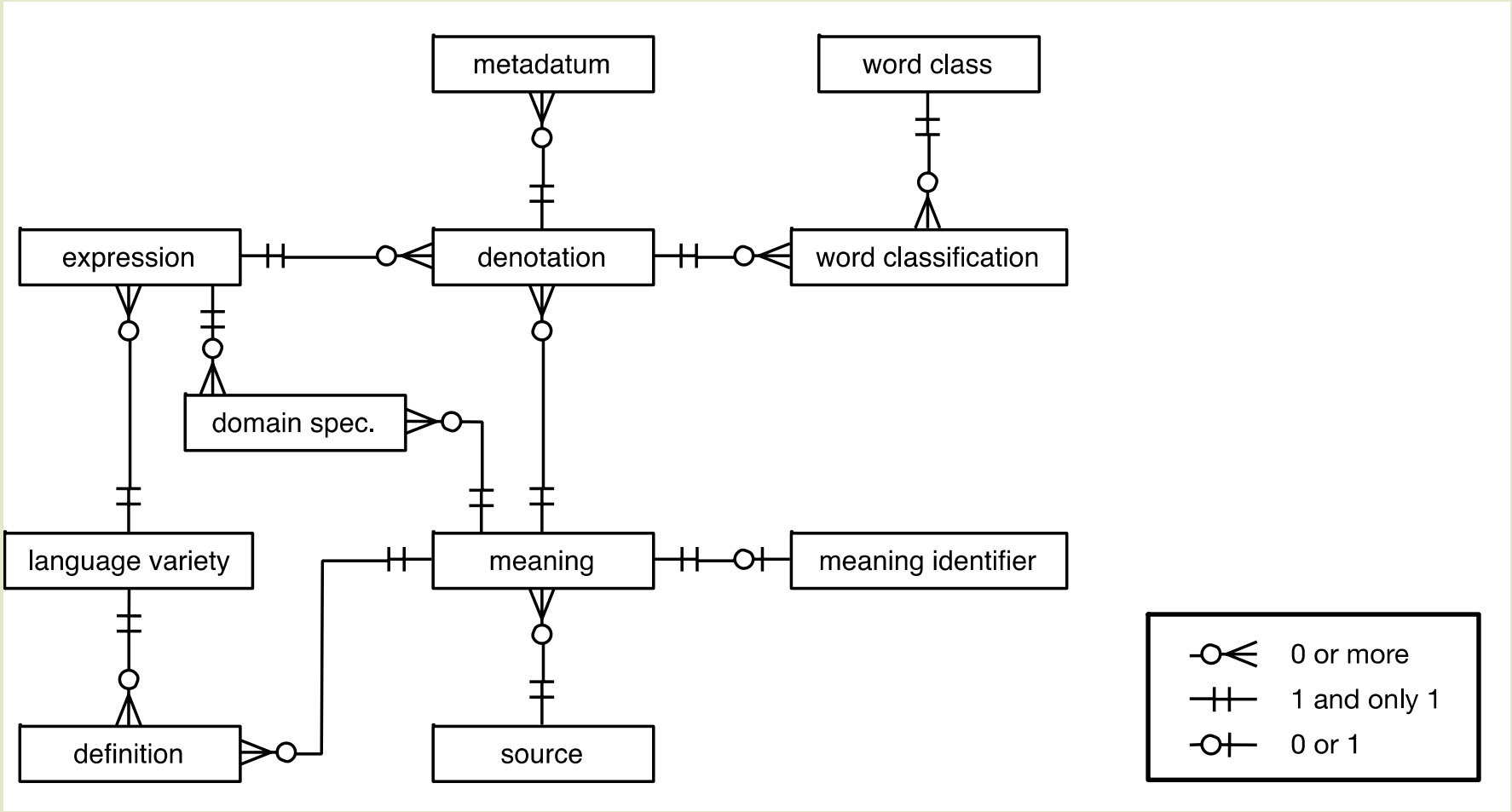
Coverage

- By number of expressions:
  - 8703 varieties with at least 2 expressions
  - 6854 varieties with at least 20 expressions
  - 2364 varieties with at least 200 expressions
  - 369 varieties with at least 2000 expressions
  - 87 varieties with at least 20,000 expressions
  - 23 varieties with at least 200,000 expressions
  - 1 variety with at least 2,000,000 expressions

Design

- Translations consist of two or more intertranslated *expressions*
  - if no translated expression is available, may have an explanatory *definition* instead
- Expression: single- or multi-word lexeme in lemmatic form
  - attributes: *lemma* (string) and *language variety*
- Language variety: identified with ISO 639 code and PanLex-specific 3-digit variety code
  - dialects and scripts treated as distinct varieties

Design: ERD


metadatum
word class
expression
denotation
word classification
domain spec.
language variety
meaning
meaning identifier
definition
source
0 or more
1 and only 1
0 or 1

Standards

- All text strings are UTF-8 NFC
- Language codes are three-char ("alpha-3") codes of ISO 639-2, 639-3, and 639-5
- Closed set of word classes is an extension of the set in OLIF
- Source documentation includes ISBN and URL

Sources

- Publicly available resources on the web
  - Wiktionaries, IDS, ASJP, freelang, many others
- Print sources
  - publicly available or purchased and scanned in cooperation with the Internet Archive
- 1500 sources have been ingested into the database so far
- 3000 unanalyzed sources in the pipeline

Source acquisition criteria

- Availability: use available publications before recruiting informants
- Tractability: prefer sources amenable to *analysis* (semi-automated computational extraction of usable data)
  - born digital sources
  - sources with consistent structures

Source acquisition criteria

- Comprehensiveness: prefer sources that document large numbers of lexical translations
- High quality: prefer works of expert lexicography to publicly edited or automatically generated sources
  - all sources rated on 0-9 estimated quality scale

Source acquisition criteria

- Coverage of low-density languages: prefer sources from languages with meager corpora and lexical documentation
- Rare language connections: prefer sources which enrich the translation graph (any-to-any rather than hub-and-spoke)

Source analysis workflow

- Tabularization: transform source data into well-defined tables with one entry per row
- Serialization: normalize data and convert tables into a format that can be ingested
- PanLex has developed various tools to manage this workflow
  - tools hosted on GitHub

Access

- Database snapshots generated monthly in CSV, JSON, and XML formats
- API provides live access to the database
  - queries and responses in JSON
- RDF interface developed at U. of Leipzig
  - data conform to lemon and GOLD data models
  - linked to Lexvo and DBpedia

License

- PanLex dataset licensed as Creative Commons CC0 1.0 Universal
- PanLex consults thousands of sources, each of which has its own copyright status or license
  - license claims are stored with individual PanLex source records
- Legal issues regarding PanLex's use of these sources not definitively answered

Applications

- TeraDict, InterVorto, TümSöz: translate lexemes from English, Esperanto, and Turkish into any other language in PanLex
- PanLinx: creates hyperlinks to expressions and translations for search engines to index
- PanLex Tattoo Generator
- Global Glossary: externally developed, based on PanLex data

Future work: 2014-2016

- Process current backlog of 3000 sources, plus about 1000 new sources, 4000 total
- About 1000 of these sources will consist of scanned images
  - OCR is impractical
  - instead, perform crowdsourced or contracted transcription and analysis

Future work: Empirical concepticons

- *Concepticon* is a recently coined term for a set of commonly translated concepts (e.g., Swadesh list, ontology)
- Most existing concepticons are curated by experts
- Alternative: derive commonly translated concepts empirically from PanLex data

Thank you!

- Project web site: panlex.org
- Source analysis workflow: dev.panlex.org
- API: api.panlex.org
- RDF interface: ld.panlex.org/rdf.html
- Database snapshots: dev.panlex.org/db/