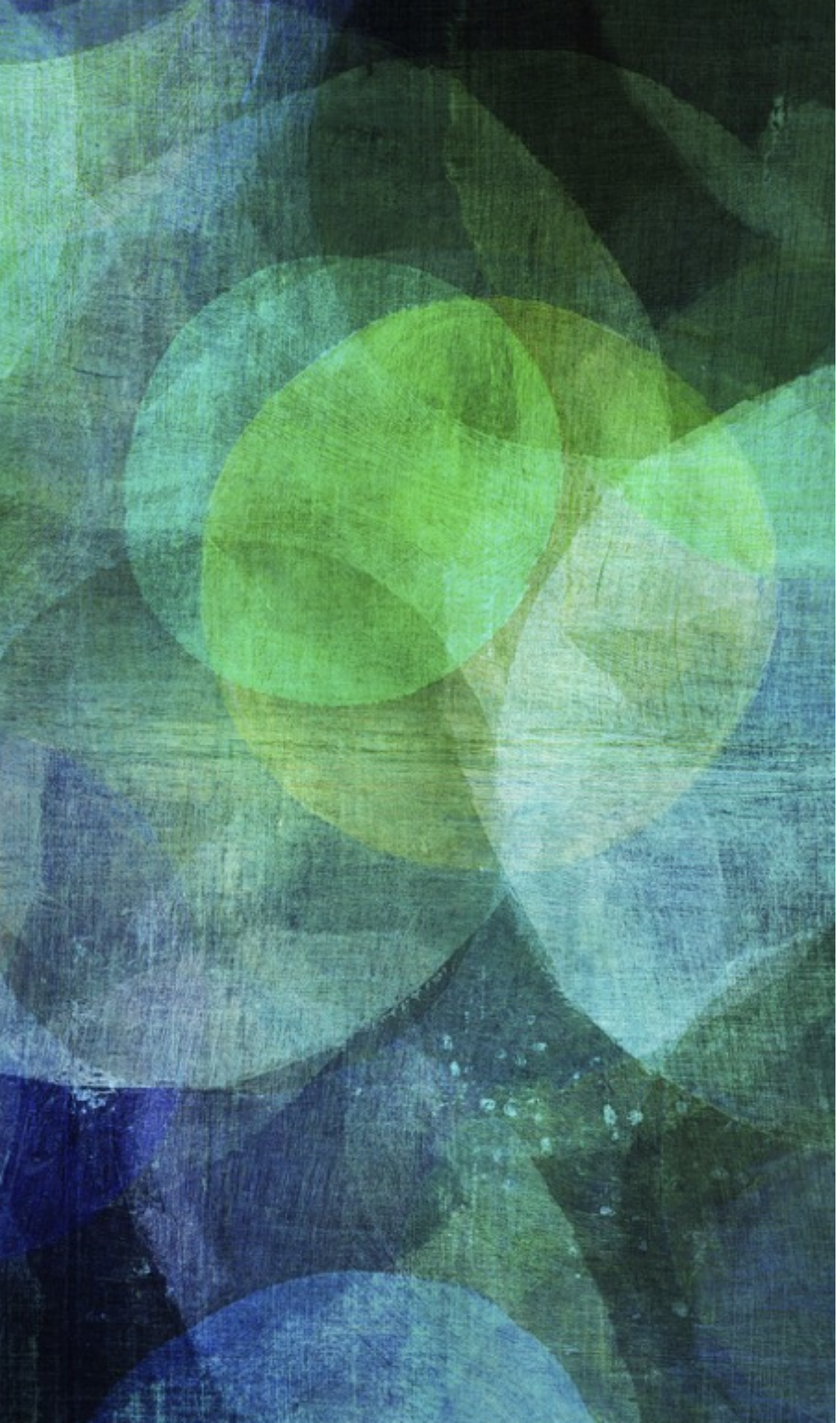
A VIEW FROM CL/NLP

Alexis Palmer
Dept. of Linguistics, Univ. of North Texas

CoRSAL Workshop
November 17, 2017



MY BACKGROUND

- PhD work: active learning for semi-automatic morpheme glossing (experimental results)
- 7 years in Germany: taught seminars on computational linguistics for low-resource languages; workshops on CL for EL documentation
- Now: at UNT, cross-lingual tool adaptation guided by linguistic knowledge; MA concentrations in Digital Language Analysis, CL

[Palmer et al. 2010, Baldridge & Palmer 2009, Palmer & Regneri 2014]

THE REALITY OF REAL DATA

- Data from several different languages, with varying degrees of relatedness
- Different stages of analysis
- Different depositors
- Different data formats
- Different approaches to annotation/interlinearization
- Different degrees of annotation/interlinearization

WHAT A MACHINE WANTS (IN SOME IDEAL MACHINE WORLD)

- Data without ambiguity
- Consistent data formats
- Consistent tag sets, consistently applied
- Structured data: each level in an interlinear text linked to other levels
- Lots of data

HARMONIZATION OF COLLECTIONS

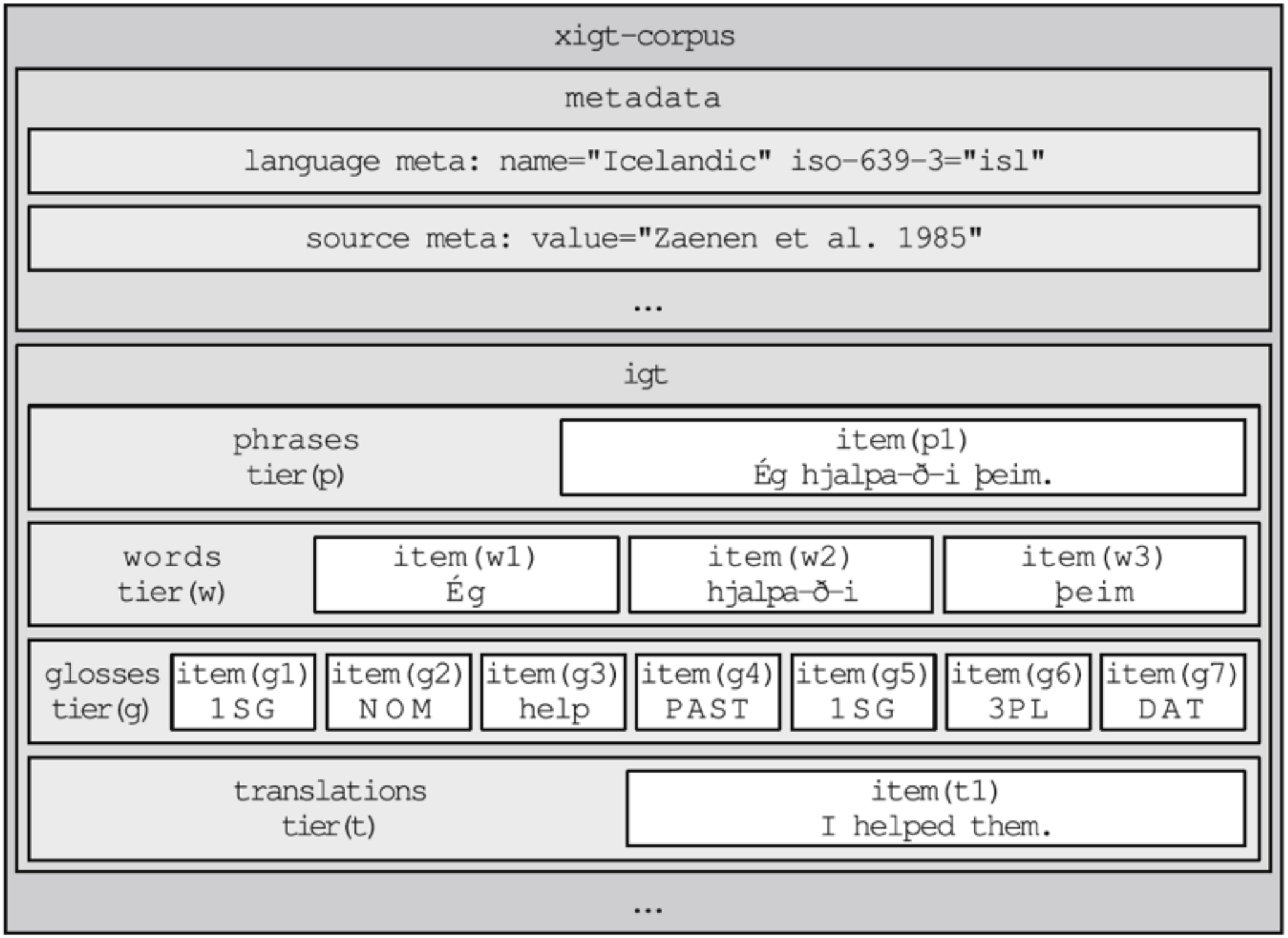
- Need: common data model for **storage** of IGT
- Once established, map data models implicit in individual collections to the common data model
- We do *not* ask depositors to put their data in this format; we may ask them to help with the mapping
- Transform collections, automating as much of the process as possible
- Various serialization formats to be considered: XML, RDF, relational DB, …
- Question: which formats best support the needs of various user groups?

XIGT (UNIVERSITY OF WASHINGTON)

- Xigt [zajkt] - Extensible representation for IGT
- Developed after detailed survey of previously-existing formats for storage of annotated language data more generally
- Discussion separates data model from serialization; in practice, data model gets realized as XML
- Advantages of Xigt
  - existing Python API - easily inspect, analyze, and modify IGT
  - tested on 1000+ languages (ODIN)
  - designed to support NLP, ML

[Goodman et al. 2015, Xia et al. 2016]

XIGT BASIC DATA MODEL – IT'S EXTENSIBLE!



CURRENT WORK IN PROGRESS

- Beta version nearly ready for release: Script to convert FLEx XML (flextext) output to Xigt
- Next: ELAN to Xigt - investigating cleanest path to lossless conversion

- Metadata at level of corpus, document, individual IGT instance (e.g. clause), individual annotation tier
- Useful in setting where any sort of automated processing is involved - detailed tracking of annotation provenance

OUTPUT (PARTIAL) OF CONVERSION SCRIPT: LAMKANG PHRASE

[http://lamkanglanguageresource.weebly.com/]

```
<xigt-corpus>
 <igt id="Bible: Psalm 23">
  <tier id="p" type="phrases">
   <item id="p1">mkpuu ngi nei'a mhai ruung chda , nei'i doo khat le kriing chni maang ; </item>
  </tier>
  <tier id="w" type="words" segmentation="p">
   <item id="w1" segmentation="p1[0:8]">mkpuu ngi</item>
   <item id="w2" segmentation="p1[10:14]">nei'a</item>
   <item id="w3" segmentation="p1[16:30]">mhai ruung chda</item>
   <item id="w4" segmentation="p1[32:32]">,</item>
   <item id="w5" segmentation="p1[34:38]">nei'i</item>
   <item id="w6" segmentation="p1[40:50]">doo khat le</item>
   <item id="w7" segmentation="p1[52:68]">kriing chni maang</item>
   <item id="w8" segmentation="p1[70:70]">;</item>
</tier>
<tier id="m" type="morphemes" segmentation="w">
 <item id="m1.1" type="prefix">m-</item>
```

EXTENDING THE VISION

- Partially-automated annotation of additional data from deposited languages ~ “bronze standard”
- Unsupervised methods - use unannotated data
- Domain adaptation methods to pivot between closely-related languages

- Scripts to transform data into desired forms (bilingual story books, for example) - made possible by common underlying data format
- Data analysis and exploration tools which learn from existing data
- ....

REFERENCES

Palmer, Alexis, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology*. 3: 1-42.

Baldridge, Jason and Alexis Palmer. 2009. How well does active learning *actually* work? Time-based evaluation of cost-reduction strategies for language documentation. In *Proceedings of EMNLP 2009*.

Palmer, Alexis and Michaela Regneri. 2014. Short-term projects, long-term benefits: Four student NLP projects for low-resource languages. In *Proceedings of Computer: The use of computational methods in the study of endangered languages*.

Wasson, C., S. Chelliah, S. Khular, and S. Basapur. 2017. Ensuring the Usefulness of a Language Archive for Indigenous Communities. Poster presented at the 2017 *International Conference of Indigenous Archives, Libraries, and Museums* (ATALM).

Goodman, Michael Wayne, Joshua Crowgey, Fei Xia and Emily M. Bender. 2015. Xigt: Extensible Interlinear Glossed Text for Natural Language Processing. *Language Resources and Evaluation*. 49(2): 455-485.

Xia, Fei, William D. Lewis, Michael Wayne Goodman, Glenn Slayden, Ryan Georgi, Joshua Crowgey and Emily M. Bender. Enriching a Massively Multilingual Database of Interlinear Glossed Text. *Language Resources and Evaluation*. 50(2): 321-349.