Harmonized Annotation and Data Formats

Discussants: Gary Simons, Helen Aristar-Dry, Alexis Palmer
Lead: Susan Kung

Definition: Harmonized Annotation

Harmonized annotation or *harmonized label sets*: (computational linguistics) compatible or interoperable formats that allow the annotations to be the same or easily translatable from one corpus/collection to another. Harmonized labels are abstract labels that annotate similarities between the different languages.

Example: corpus X uses the annotation term ‘perfective’, corpus Y uses ‘past tense’, corpus Z uses ‘completed’. Harmonized annotation might include a tier that labels all of these annotations as the same thing (e.g. perfective).*

* Setting aside distinctions between tense and aspect for the sake of a quick and easy example.

Question 1: What issues exist that require concerted efforts towards harmonized annotation

Harmonized annotation will make data comparison and analysis easier in areas such as

- Computational linguistics and natural language processing
  - Many models learn from data
  - More instances of fewer labels = greater capacity for generalization
- Typological/lexical/etc. analysis
  - Within a language family and beyond
  - Areal analysis

Question 1a: How can harmonized annotation facilitate cross-linguistic analysis?

- E-MELD (Electronic Metastructure for Endangered Languages Data)
- GOLD (General Ontology for Linguistic Description)
- Coarse-grained label sets, e.g. Universal POS Tagset (POS=part of speech)
- Mapping depositor's tagsets to the CoRSAL tagsets
- IGT = Interlinear glossed text

Question 1b: What are some current hurdles to achieving harmonized data across all data in an archive?

- Determining the label sets
- Linguists
  - Strong feelings about their labels & more than 1 linguist = disagreement
  - Standardized orthography debates
- Software development
  - Helen Aristar-Dry: "Ask less of the linguist and more of the software."
- Application of the harmonized labels
  - CoRSAL staff vs. depositor/linguist (see above)
  - Training
- Quality Control or Assurance after application of labels

Question 1c: What are some ***future*** hurdles to sustaining harmonized data across all data in an archive?

- What happens when an archive user wants to ask new questions of the corpus that cannot be answered with the existing label sets? Can they make a request to CoRSAL to apply a new label set?
- Can anyone make the request or just computational linguists who will know how to write the script that will apply the new label set? Will there be a CoRSAL staff person who will write the script?
- Version control (before and after application of new label)
- QC/QA for second round of labeling.

Question 1d: What types of data or speech events would be appropriate for harmonization?

- Stories, myths or folktales
- Oral History
- Conversation
- Oratory events, prayers, speeches
- Songs, chants
- Grammatical elicitation
- Word lists or lexical elicitation
- ...

Question 2: What solutions should be sought to reach harmonized annotation?

- Determination or standardization of the harmonized label sets.
  - Plan for how to add label sets in the future.
- Automated machine harmonization
  - Will this require a language model for every language? Just every language family?
- Copyright associated with permission to create derivative data
  - Carefully consider all aspects of copyright around these data, including Intellectual Property rights of the ***speakers*** and the Traditional Knowledge rights of the ***community.***

Definition: Data Formats

- **Media type**: audio, video, images, text, binary, and PDF/A files (each media type will require an online viewer (most complex for text and binary media))
- **File formats**: (digital archives typically avoid including proprietary formats)
  - Audio: .wav, .mp3
  - Video: .mpg, .avi, .mov, .mp4, .m4v
  - Image: .tif, .tiff, .jpg, .png
  - Text/binary: .txt, .xml, .flextext, .eaf, .trs, .csv, .rdf, .html, etc.
  - PDF/A (archival standard): .pdf
- **Structure of the file contents**: how the data is formatted within the file
  - Interlinearized and glossed text (Elan, Transcriber, etc.)
  - MS Word-like formatting with changes in font and color
  - Databases, e.g. (domain-specific) dictionaries
  - Hyperlinked documents

Question 3: Why are comparable data formats desirable?

- Value of transforming various input data formats into one common xml-based format (e.g. Xigt, work out of U Washington)
  - Clear underlying data model
  - Extensible: easily accommodates files with different amounts of analysis/annota none!)
  - Serialization is separate from data model: can be transformed into (e.g.) RDF, re database, JSON, etc.
  - Designed specifically for language data
  - Existing import/export scripts: Toolbox, FLEx
- Benefits of making IGT machine-readable
  - Rich source of linguistic information
  - Train systems to support future linguistic analysis - speed up the process!

Question 4: What are the practical roadblocks and theoretical problems in creating comparable data formats?

- Not all IGT is machine-readable (PDF, WORD, handwritten manuscripts)
- Challenges of harmonized annotations/tag sets

Question & Further Discussion

...

Thank you!