ANNOTATION Overview

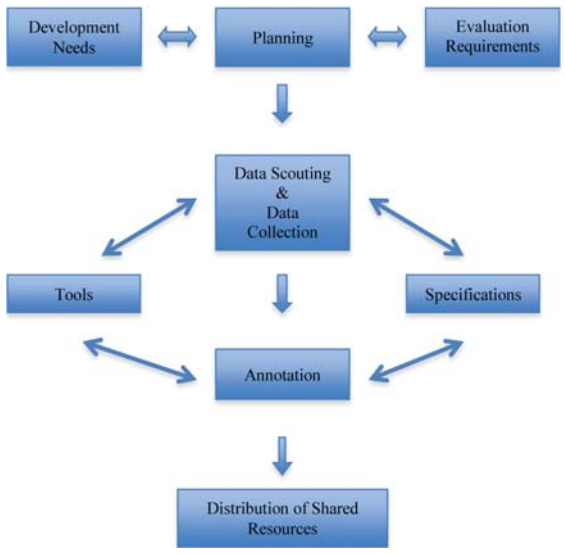
What is Annotation?

Annotation describes any direct or computer-mediated application of human judgment, description or analysis that adds value to raw linguistic data. Annotation encompasses a wide variety of task types including, but not limited to, transcription, translation, segmentation, tagging, parsing, analysis and disambiguation. Raw data may be in audio, video or text forms.

Language research and technology development require large volumes of raw data processed into a usable form and enhanced by annotations of an appropriate type, richness and quality. Annotations support many research projects and human language technologies, such as speaker recognition, speech-to-text and text-to-speech, machine translation and information retrieval and extraction

LDC Annotation Workflow

Creating annotated data relies on a robust pipeline, trained annotators, clear annotation specifications and documentation, consistent file forms and appropriate software. LDC regularly develops infrastructure in each of these areas. The chart below represents a typical workflow of a project from planning stages through the publication of resources.



Typical annotation workflow

Types of Annotation and Lexicography

Speech

- Time-aligned orthographic transcription
- Story, speaker turn, word segmentation and alignment
- Speaker and language identification and recognition
- Discourse structure and disfluency labeling
- Prosodic marking
- Sociolinguistic variable coding

Translation

- Translation, multiple translation and quality assurance
- Document, sentence, phrase and word level parallel text alignment
- Post-editing and evaluation of MT output

Syntactic, Morphological, Sematic

- Treebanking, Propbanking
- MPG (morphological, part-of-speech, gloss) tagging
- Identification and classification of entities, relations, events, temporal expressions and co-reference
- Entailment tagging and knowledge base population
- Sense disambiguation

Image and Video

- Video entity and event labeling and co-reference
- Video and image zoning and text transcription
- Ground truthing at line, word and sub-word levels
- Handwritten document classification, legibility and readability
- Human gesture labeling

Lexicons

- Traditional
- Pronunciation with morphology
- Translation

Document

- Fine and coarse grained topic labeling
- Novel information
- Single and multi-document summarization
- Mono and multilingual summarization

Linguistic Data Consortium 3600 Market St., Suite 810, University of Pennsylvania, Philadelphia, PA 19104-2653 USA
Telephone: +1.215.898.0464 • Fax: +1.215.573.2175 • ldc@ldc.upenn.edu • www.ldc.upenn.edu

Annotation Tools

To maximize the efficiency of annotation teams, LDC creates new processes and customizes interfaces that both drive annotation tasks and exploit existing annotations. LDC tests, process engineers and develops specifications that simplify the annotation task and reduce cognitive load in order to increase speed and consistency and reduce errors. Examples include:

- XTrans Speech Annotation Tool: A multi-platform, multilingual, multi-channel transcription tool developed to support manual transcription and annotation of audio recordings. **http://www.ldc.upenn.edu/tools/XTrans/**
- Champollion Toolkit (CTK): Built around LDC's Champollion sentence aligner kernel, CTK provides ready-to-use parallel text sentence alignment tools for many language pairs. **http://champollion.sourceforge.net/**
- Arabic Reading Enhancement Tools: A series of tools designed for Arabic reading facilitation and assessment, morphological analysis, dictionary lookup and word concordance output. **http://web1.ldc.upenn.edu/Projects/art/**

LDC word-alignment tool

- LDC Word Aligner: A tool for manual annotation of word alignment relationships between a text and its translation in a different language. **http://www.ldc.upenn.edu/tools/WA/**

Annotation Specifications

LDC develops specifications for all of its annotation tasks building on existing guidelines and published papers or, when necessary, creating new specifications for novel tasks through needs assessment, experimentation, measuring inter-annotator agreement and evaluating the resulting annotations for downstream purposes.

Annotator Training and Quality Assurance

Annotator training is the first step in ensuring quality. LDC selects, tests, vets and trains annotators for tasks and assignments appropriate to their skills and abilities. Annotators in training read guidelines, observe demonstrations, practice annotations and participate in group meetings, face-to-face and computer-mediated discussions. LDC further assures quality through the use of specialized interfaces that guide annotation. Other measures include multiple annotation passes and annotator testing against reference data sets both to measure inter-annotator agreement and to identify areas that need further attention.

To measure agreement a subset of all decisions are assigned to multiple annotators using a double blind methodology that eliminates the possibility that either the annotators being compared or their supervisor knows when the testing is taking place. Once initial agreement rates are known, LDC managers monitor the impact of changes in tasking with an eye toward maximizing consistency.

Best Practices

The use of publically accessible standards improves the exploitation and enables the sharing of Language Resources (LR). LDC adopts existing best practices or, where absent, establishes new ones for LR creation and dissemination. LDC documents its practices through annotation and format specifications, which are typically made publically available in conference presentations, published papers and on the LDC website.

Annotation Sponsored Projects

LDC has supported numerous programs actively working with sponsors to plan, develop, test, create, annotate and distribute linguistic corpora, including RATS, BOLT, HAVIC, MADCAT, GALE, TIDES, EARS, ACE, TalkBank, MR, HARD, TRANSTAC, VACE, AQUAINT, TDT and TREC, as well as many NIST technology evaluation campaigns.