Managing a Multilingual Treebank Project

Milan Souček, Timo Järvinen, Adam LaMontagne
milan.soucek@lionbridge.com, timo.jarvinen@lionbridge.com, adam.lamontagne@lionbridge.com

Challenges

- Simultaneous treebank development in multiple languages
- High quality requirements
- Cross-language consistency requirements
- Multilingual, distributed team of language experts
- Controlled budget
- Strict deadlines

Achieving consistency

Technical prerequisites

- Work environment (Cloud)
- Tools (TrEd, SVN)
- XML schema validation
- QA validation tool

Annotation consistency

- Hands-on trainings for annotators
- Team discussion (online discussion board)
- Annotation cross-check – annotators checking each other's results
- Lead linguist – reviewer model
- Cross-language feedback – lead linguists review other language's data to agree on consistent cross-language model

Process flow

➢ Process cycle consists of

1. Data parsing
2. Manual annotation and review
3. Three-level validation and
4. Parser training

✓ Iterative parser training improves annotation efficiency and throughput

✓ Validation assures consistent output

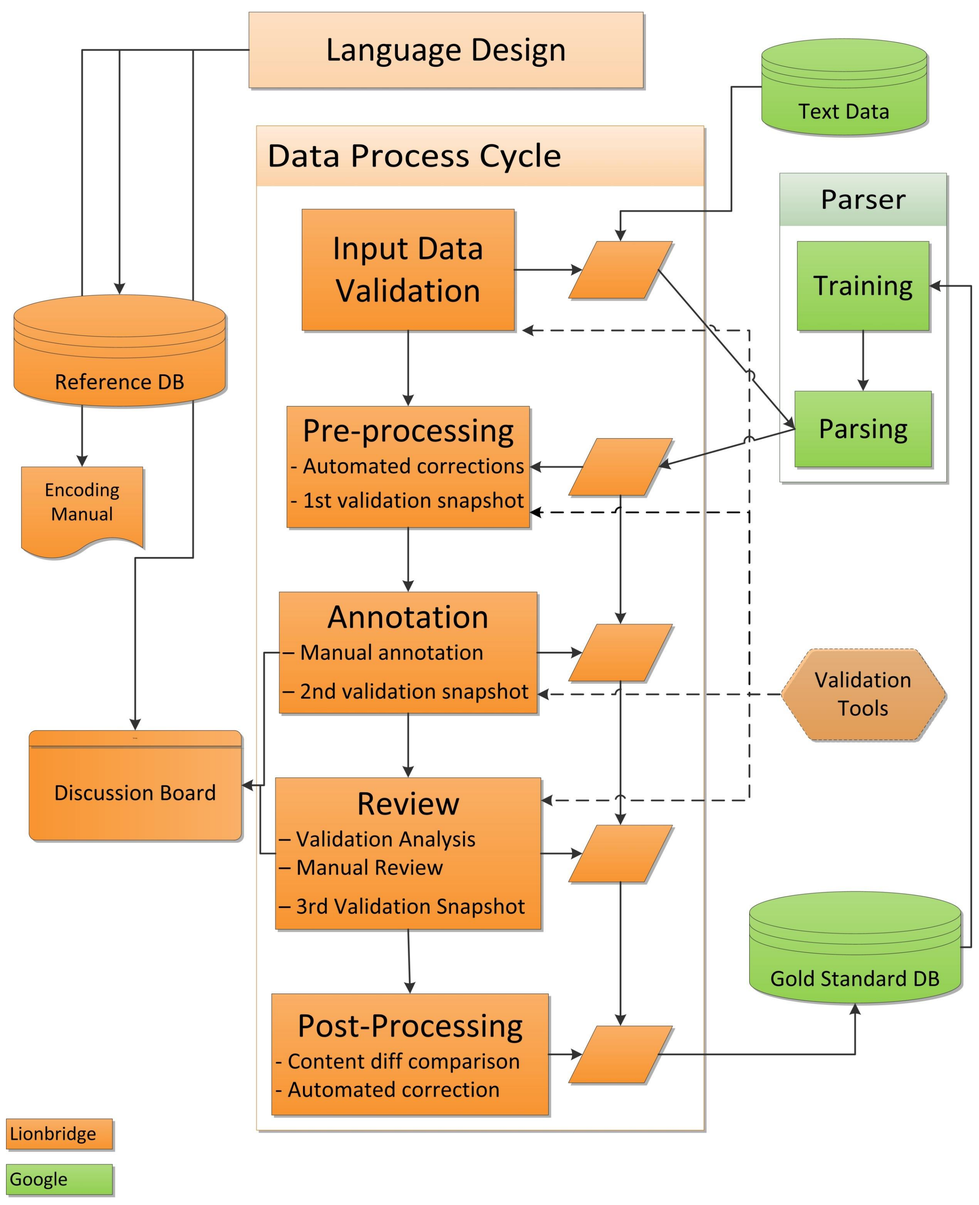

Figure 1 – Process cycle flowchart

Validation

- **On-line**: lists of valid POS and deprel labels in the XML schema file
- **Semi-automated post-validation**: POS vs. deprel representation check based on predefined possible/frequent POS combination for deprel participants

Dependency label	parent POS	dependent POS
advcl	VERB	VERB
advmod	VERB, ADJ, ADV	ADV
amod	NOUN, X, PNOUN	ADJ

Figure 2 – Example of validation tool settings

S-ID	porder	order	form	deprel	postag	ppostag	Validate-postag	Validate-ppostag
s-1	5	9	angebracht	scomp	ADJ	VERB	Valid	Valid
s-1	9	6	in	prep	ADP	ADJ	Valid	Valid
s-1	9	8	Einheitsdrehgestell	pobj	NOUN	ADJ	Valid	Wrong
s-1	6	7	einem	det	DET	ADP	Valid	Wrong

Figure 3 – Example of validation tool output

Results achievements

➢ Dependency treebanks developed in four languages during the first phase of the project:

Treebank model: Stanford typed dependency

Languages: French, German, Spanish, Brazilian Portuguese

Volumes: 15k sentences per language (Wikipedia and news data)

Completion time: 6 months

Annotation throughput: Initial throughput between 8-12 sentences per hour, improved to 30-40 sentences at the late stage of the project

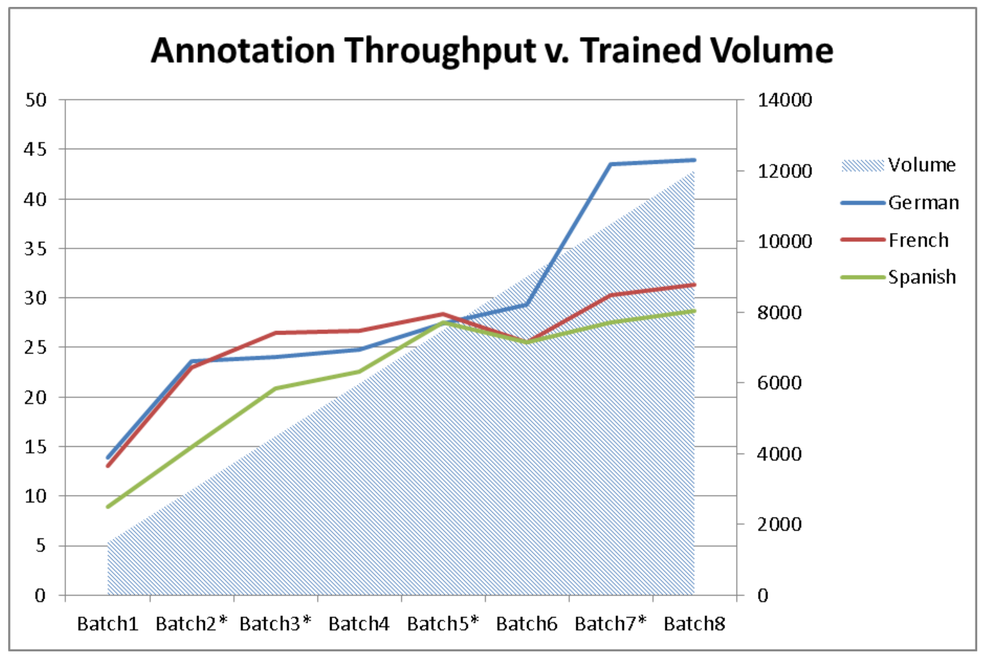

Figure 4 – Annotation throughput vs. trained volume projection

Future development

- Further languages being added
- Using experience from pilot languages for creating consistent multilingual set of treebanks
- Developing more sophisticated validation methods
- Experiments with treebank conversion
- Research on linguistics universals for syntactic parsing

Acknowledgments

- Thanks to the team at Google for giving us the opportunity to work on this fascinating project.
- Thanks also to Lionbridge for providing the environment, means and support to conduct this research
- Special thanks to all our language teams for their hard work and dedication to making this project a success

Lionbridge