Today's Proposed Schedule

08:00 – Introductions
08:20 -- ARL MURI vision -- Joe Meyers
08:40 -- Overview of Linguistic Core MT -- Jaime
09:00 – Language Selection for MT -- Jason
09:30 -- Comparative Linguistics -- Lori
10:00 -- Break
10:30 – Multilingual Language Learning -- Regina
11:00 – Combining Data and Linguistics -- Noah
11:30 -- Issues to discuss/resolve -- All
12:00 -- Lunch
12:30 -- Evaluation -- Stephan
13:00 -- Milestones/ Management -- Jaime
13:30 -- Government huddle
14:15 – Government report
14:30 -- General discussion
15:00 – Meeting concludes

The Linguistic-Core Approach to Structured Translation and Analysis of Low-Resource Languages

Kickoff Meeting for ARL MURI Project

26-October-2010

The Cast

CMU:

Jaime Carbonell 

 Lori Levin

Stephan Vogel 

 Noah Smith

ISI:

Kevin Knight 

 David Chiang

MIT:

Regina Barzilay 

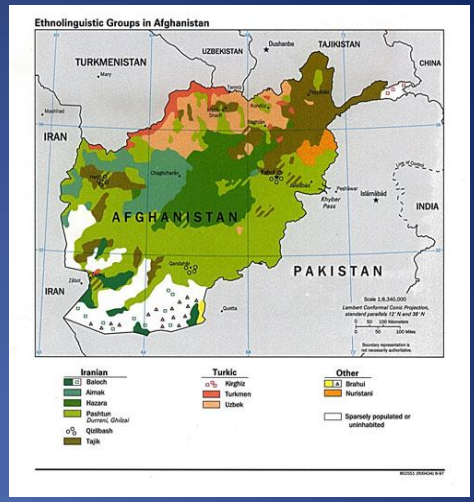
UT:

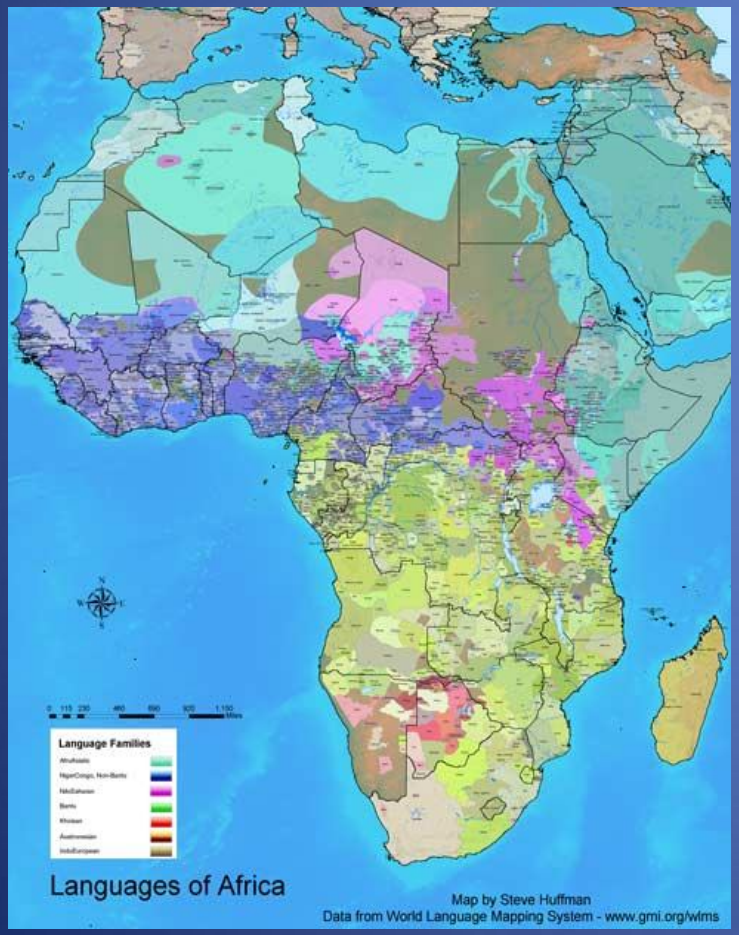
 Jason Bladridge

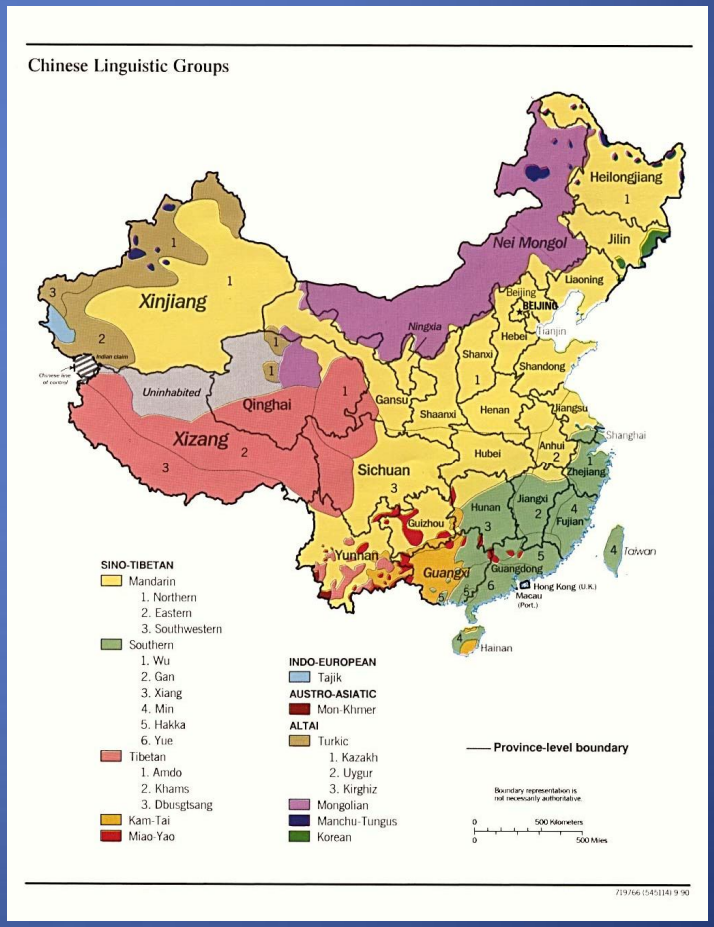
Supporting roles: Graduate Students, Postdocs, Informants, …

The Setting

6K languages → 3K with 10,000 speakers







The Setting

- LR Languages, e.g. in Africa, cannot be ignored
- MT & TA for LRLs requires a linguistic core
  - Insufficient parallel text for standard SMT
  - Insufficient annotations for purely statistical TA
- Phrasal SMT, even for HRL, errs
  - E.g. divergences, long-distance movements,...
- But Computational Linguists are Expensive
  - Cannot dedicate person-centuries per language to write, test and debug massive rule-based systems
  - Army needs a more rapid & cost effective approach

The Scientific Questions

- Can deep linguistic representations benefit practical MT & TA?
- Can we marry learning from data with expert-crafted declarative linguistics?
- Can we uncover underlying linguistic structure through comparative language analysis?
- How can we extend MT-motivated linguistic-core capabilities to related TA tasks?
- Can different linguistic analyses reinforce each other synergistically?
- How important is resolving complex morphology?
- How important are general semantic features for MT?
- How well can unsupervised learning methods augment linguistically motivated analyses for MT and TA?
- ….

Which MT Paradigms are Best? Towards Filling the Table

Target

Source

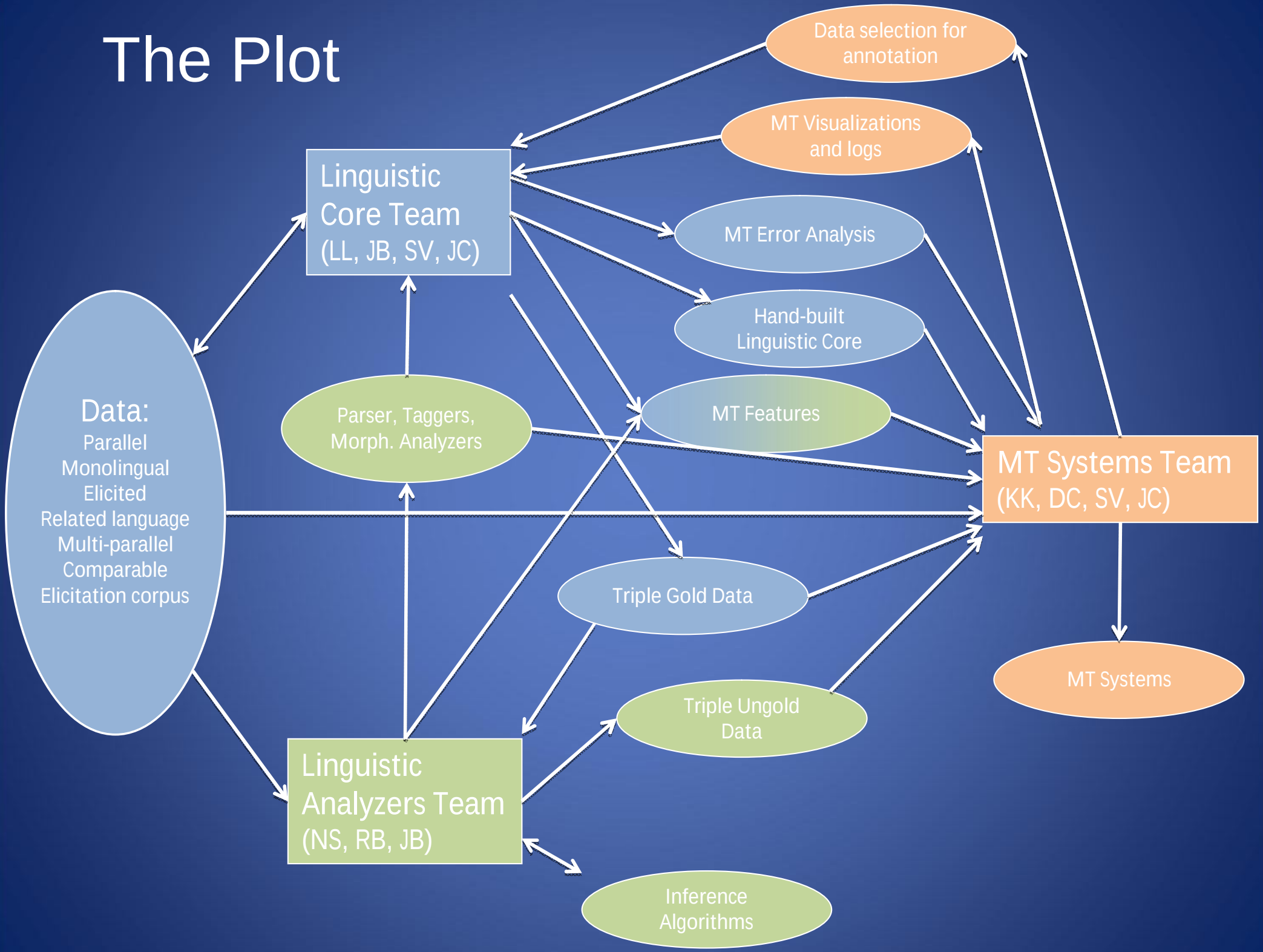
	Large T	Med T	Small T
Large S	SMT →	???	???
Med S	??? ↓	???	???
Small S	???	???	???

- DARPA MT: Large S → Large T
  - *Arabic → English; Chinese → English*

The Approach

- Linguistic core: Universals & Specifics
  - Specialize core to each language pair
  - Active Learning from elicitation corpus
- Learning to extend the core
  - Supervised active learning
  - Unsupervised from monolingual & comparable corpora & related languages
- Exploration of multiple paradigms
  - E.g. Dependency parsing
  - E.g. Finite-state transducers
  - Ensemble methods
- Build, evaluate, refine glass-box end-to-end prototypes

The Plot


Data selection for annotation
MT Visualizations and logs
Linguistic Core Team (LL, JB, SV, JC)
MT Error Analysis
Hand-built Linguistic Core
Data:
Parallel
Monolingual
Elicited
Related language
Multi-parallel
Comparable
Elicitation corpus
Parser, Taggers, Morph. Analyzers
MT Features
MT Systems Team (KK, DC, SV, JC)
Triple Gold Data
MT Systems
Triple Ungold Data
Linguistic Analyzers Team (NS, RB, JB)
Inference Algorithms

THANK YOU!