The Matrix: Future Directions
Wrap up

Ling 567
March 5, 2012

Overview

- Wrap up/reflections
- Matrix: Future directions
    - More libraries
    - More robust MMT
    - Applications, including language documentation

Goals: Of Grammar Engineering

- Build useful, usable resources
- Test linguistic hypotheses
- Represent grammaticality/minimize ambiguity
- Build modular systems: maintenance, reuse

Goals: Of this course

- Mastery of tfs formalism
- Hands-on experience with grammar engineering
- A different perspective on natural language syntax
- Practice building (and debugging!) extensible system
- Contribute to on-going research in multilingual grammar engineering

Reflections

- Where have the analyses provided by the Matrix (or suggested by the labs) seemed like a good fit?
- Where have they been awkward?
- What have you learned in this class about syntax?
- ... about knowledge engineering for NLP?
- ... about computational linguistics in general?
- ... about linguistics in general?
- What did working with a test corpus show you about the process of scaling to real-world text?

Feedback: Pair projects

- How did you divide the work?
- In what ways was having a partner helpful?
- Would you have learned more working on your own?

Future directions overview

- More libraries (and semantic harmonization)
- How this class will evolve
- MT: Auto-generated transfer rules, typological seeding of statistical NLP (including SMT)
- Lexical acquisition
- Ontological annotation
- Matrix-ODIN Mash-up

More libraries

- New this year: Updates to negation
- Demonstratives
- Extensions/retrofits questions, coordination
- (more) extensions to word order
- Non-verbal predicates
- Intersective modifiers
- Numeral classifiers
- More verb subcategorization
- Embedded clauses
- Marking of information structure

More libraries/reflection from current class

- What do you most wish was available in the customization system, based on what came up in your test suite?
- In your test corpus?
  - Adj/adv; multiple inheritance of lexical types; sentence fragments; free affix order; medium flexibility word order, e.g. V-final and V2;
  - Split ergativity: can’t just talk about tense X, had to do trans and intrans separately; had to define one extraneous case (abs)
  - Import two choices files and output a documented diff
  - Class feedback: Rather than chain of identities in lab 1, take that English grammar and make a small modification, especially one that illustrates how lex rules work in tdl.

Evolution of 567

- New phenomena: Wh-questions, relative clauses, while-clauses ...?
- Ever bigger jump start --- reaching the limit on this one?
    - Would working in groups of three make it possible to get to even bigger grammar fragments?
- New this year/how did these work out?:
    - Partnership with field linguists
    - Work with small corpora
- Coverage-driven labs seem most satisfying (MT demo, corpus coverage). Is this true? Can the course be rebalanced to do more of this?

Lexical acquisition

- How can we import lexical entries from other linguistic resources (e.g., FIELD lexicons, ODIN)?

- How big do the grammars have to get before we can embark on (semi-)automated lexical acquisition?

- To what extent do the lexical properties of translational equivalents predict lexical properties in another language?

- How can we most effectively leverage human effort?

- How do we know when we're missing an appropriate type?

Autogenerated transfer rules

- Identify “grammaticized” differences in MRSs
- “Publish” choices along these dimensions for each grammar
- Create a library of transfer rules from property to property:
  - pro-drop to pronouns (and vice versa)
  - mismatches in demonstrative distinctions
  - can <> the possibility exists
  - hurt/cause feel+pain/cause harm

Autogenerated transfer rules

- Use language-specific pred values
- Create transfer rules on the basis of PanDictionary or other lexical resources
- Measure the extent of translation divergence (Francesca Gola's MS thesis work)
- Use bitexts and statistical methods to detect word pairs requiring more than straight pred-mapping transfer rules

Seeding statistical NLP with typological knowledge

- Haghighi & Klein 2006: Unsupervised parsing (“Prototype Driven Grammar Induction”)

- Syntax-based statistical MT is finally coming into its own (e.g., work by Callison-Burch)

- Matrix Customization system-generated starter grammars represent a middle ground between broad-coverage precision grammars and coarse-grained typological information (as in WALS).

  - Testable over hand-constructed test suites

  - Usable to create prototype trees or even translation pairs

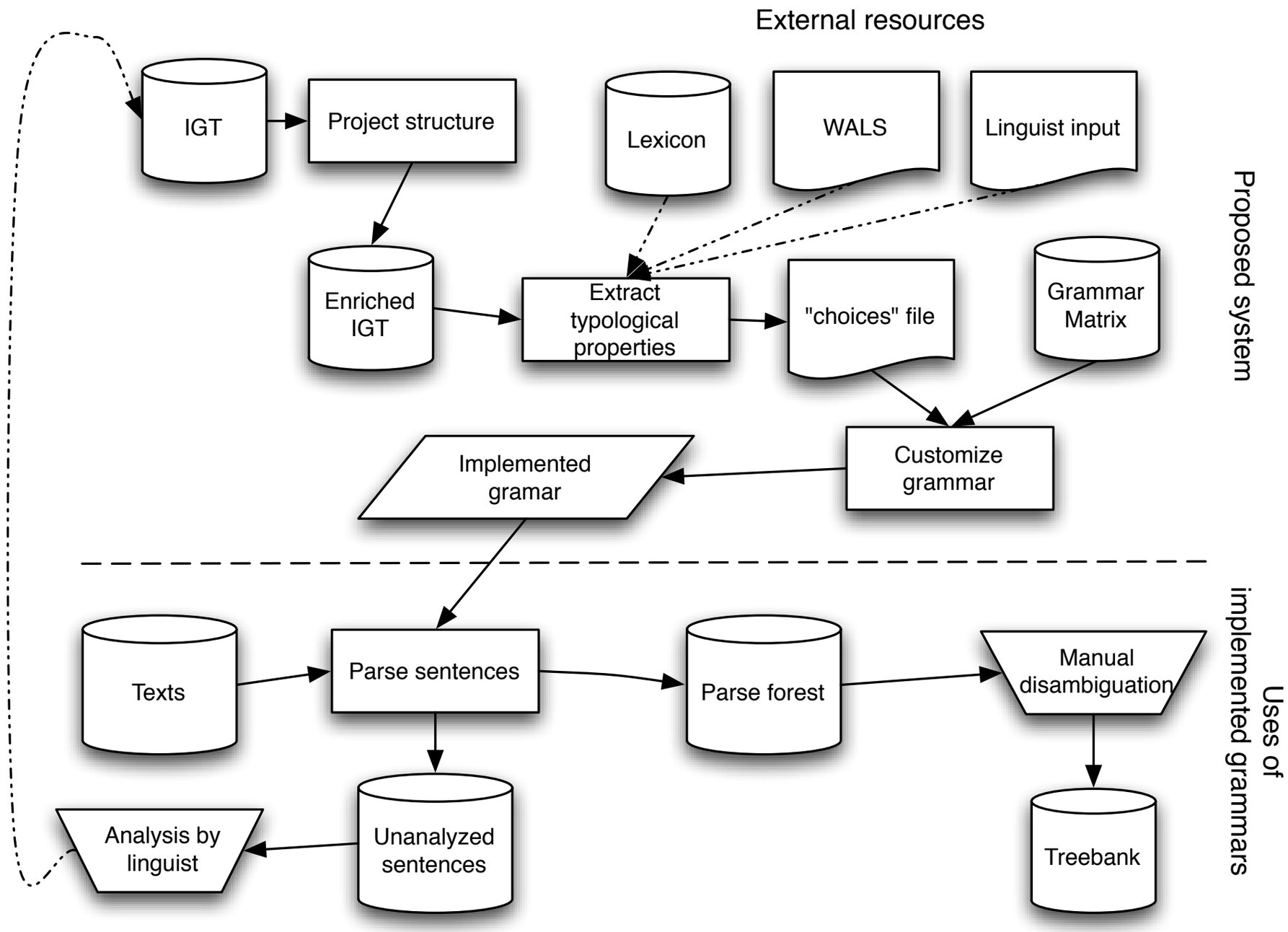
Ontological annotation

- Annotate grammars with links to GOLD (Farrar & Langendoen 2003)
    - Locate which constraints contribute to which phenomena
    - Index analyses for discovery in grammars and treebanks
- Annotations in Matrix core
- Annotations in customization system
- Support for user annotation

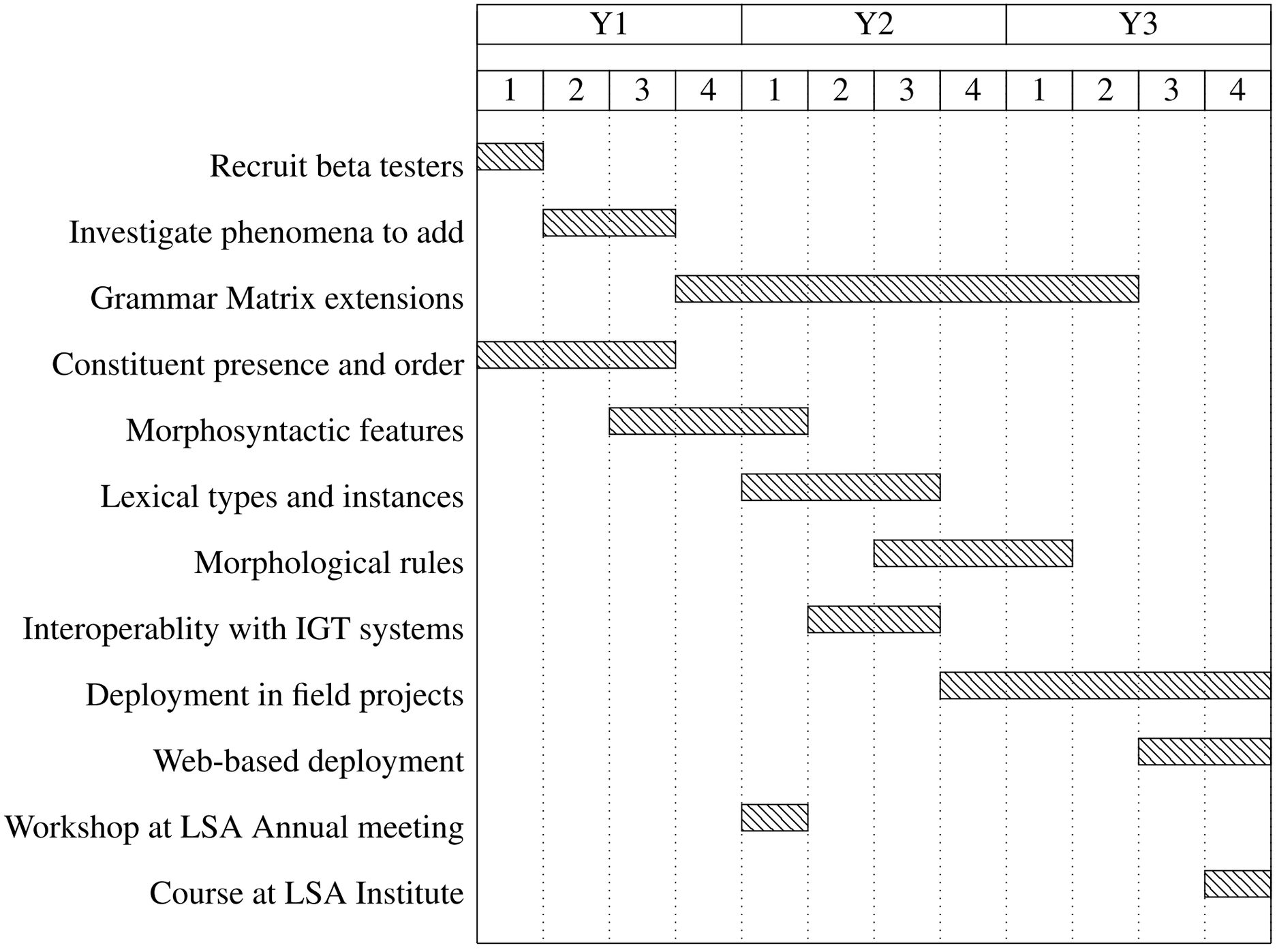
Matrix-ODIN Mash-up

- ODIN: Online Database of INterlinear glossed text (Lewis 2006)
- Lewis & Xia 2007 explore learning typological properties from ODIN data
- Next steps:
  - Answer Matrix customization system questionnaire automatically
  - Including lexical information
  - ... and information about affixes (David Wax's MS thesis work)
- Step 3: Automatically generated precision grammars!

MOM for Documenting Endangered Languages (NSF DEL proposal under review)



MOM for Documenting Endangered Languages (NSF DEL proposal under review)



Overview

- Wrap up/reflections
- Matrix: Future directions
    - More libraries
    - More robust MMT
    - Applications, including language documentation