





Miriam Butt University of Konstanz

### Course Content

In this course you will learn:

- How to implement a computational grammar fragment within LFG (Lexical-Functional Grammar).
- Work with the grammar development platform XLE.
- Learn how to integrate finite-state morphological analyzers into the grammar.
- At the end of the course you should be able to begin building your own grammar for a language of your choice.

#### Course Content

What you will **not** learn:

- The course will introduce basic concepts and standard analyses within LFG.
- But the course will spend no time motivating these analyses or delving into alternatives.
- You will need to take a course on LFG (also available on-line) or work through the relevant literature (provided as part of the course).

#### Course Structure

- See the accompanying file for a detailed course outline.
- Each lesson contains:
  - I) introduction of the theoretical concepts
  - 2) demonstrations of how to implement the concepts
  - 3) a sample grammar fragment
  - 4) exercises working with the grammar fragment

### **Course Topics**

- I. LFG Basics and First Steps in XLE (Walkthrough)
- 2. Lexical Rules: Passives and other Argument Alternations
- 3. Adjuncts: Adverbs and Adjectives (also xlerc file)
- 4. PPs: Adjuncts and Obliques
- 5. Generation and the Optimality Projection
- 6. Complements: XCOMP and COMP
- 7. Long Distance Dependencies and Functional Uncertainty

## Course Topics II

- 8. Pronouns, Empty Nodes and Punctuation
- 9. Integrating a Finite-State Morphological Analyzer
- **10.** Further Basic XLE Features
  - Shuffle Operator (Free Word Order)
  - Variables (Relative Clauses)
- II. Outlook: Advanced features of XLE
  - Integration of Statistical Preferences
  - Complex Categories, Transfer System, Restriction Operator

#### Outline, Lesson 1

- I. Motivation: Why Deep Grammars?
- 2. Demos of possible Applications
- 3. XLE/LFG Basics
- 4. The ParGram (Parallel Grammars) Collaboration

# Why Deep Grammars?

- I. Some Background and Motivation
- 2. Examples of Applications
  - Question-Answering Systems
  - Electronic Look-up/Learning Resources
    - Computer Assisted Language Learning (CALL)
    - Murrinh-Patha English Translation System
  - Automatically Annotated Corpora (Treebanks)

## Deep Grammars

- Provide detailed morphosyntactic information
  - Include detailed information about dependencies within the clause (subject, object, clausal complements).
  - Tense/Aspect/Mood
  - Person/Number/Gender
  - Modification, Specification (Definite/Indefinite)
- LFG also provides detailed information about constituency and hierarchical relations of words in the clause.

### Deep Grammars

• Simplified Example:

Mary wants to leave.

subj(want~1,Mary~3)
comp(want~1,leave~2)
subj(leave~2,Mary~3)
tense(leave~2,present)

- Deep Grammars are usually manually constructed.
- Large Multilingual Grammar Development Efforts include:
  - LFG (ParGram)
  - HPSG (LinGO, Matrix, DELPH-IN)

## **Applications:**Treebanks

- Many Natural Language Processing (NLP) tools depend on linguistically sophisticated resources.
- One example of such a resource is a so-called treebank.
- A treebank or *structure bank* is a corpus of sentences annotated with deep linguistic information.
- Deep grammars can be used to automatically annotate such a corpus of sentences.
- This also serves to help with *language documentation*.

## Applications: Q&A

- Deep grammars also have potential to be used in sophisticated Question-Answer (Q&A) systems.
- Examples:
  - IBM's Watson (uses a dependency grammar)
  - The PARC Bridge System (led to founding of the company Powerset)

#### IBM's Watson

IBM's Watson made a spectacular debut playing against humans on the US game show Jeopardy.



#### IBM's Watson

- Check it out on YouTube.
- Read about the design here (and elsewhere):

Ferrucci, David, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer & Chris Welty. 2012. Building Watson: An overview of the DeepQA Project. *Al Magazine* 31(3).

- Researchers at the Palo Alto Research Center (PARC) put together a demo to show-case their ideas for a linguistically sophisticated Q&A system.
- The system includes
  - LFG grammars of the type constructed in this course.
  - Finite-State Morphological Analyzer
  - An Abstract Knowledge Representation System (AKR)
  - Integration of lexical semantic information from WordNet and other lexical resources/ontologies.

Read about the system here:

Bobrow, Danny G., Cleo Condoravdi, Kyle Richardson, Richard Waldinger & Amar Das. 2011. Deducing answers to English questions from structured data. 2011. International Conference on Intelligent User Interfaces (IUI). Stanford.

Bobrow, Danny G., Robert D. Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy H. King, Rowan Nairn, Valeria de Paiva, Lotti Price & Annie Zaenen. 2007. PARC's Bridge question answering system. Proceedings of the GEAF (Grammar Engineering Across Frameworks) 2007 Workshop. Stanford.

00	X parc bridge
	Display Analyses: 🔶 single 💊 packed 🛛 Clear 🛛 Exit 🛛 Help
Passage	A man vanished last Saturday.
(	Syntax AKR0 AKR
Question	
Question	Did a person disappear?
	Syntax AKR0 AKR
Answe	r YES: [person=man,disappear=vanish]

## Underlying LFG Grammar



Each sentence is parsed by a large LFG grammar for English developed at PARC.

Each parse contains a constituent structure (c-str).

This shows linear order of the words, constituency and the hierarchical organization of the constituents.



A finite-state morphological analyzer (FSM) feeds into the grammar.



The functional structure (f-str) provides dependency and other deep linguistic information.

00	X parc bridge
	Display Analyses: 🔶 single 🔷 packed 🛛 Clear 🛛 Exit 🛛 Help
Passage	A man vanished last Saturday.
	Syntax AKR0 AKR
Question	Did a person disappear?
(	Syntax AKR0 AKR
Answe	er YES: [person=man,disappear=vanish]



## Semantics (AKR)

- The f-structure provides the input for the semantic analysis, the Abstract Knowledge Representation (AKR).
- F-structures have been shown to be equivalent to Quasi-Logical Forms (QLFs) so they already contain a good approximation of basic semantics.
- The AKRs integrate lexical semantic knowledge from WordNet and other sources.
- The AKRs of the two sentences are compared to see if the answer to the question can be judged as correct

ill	
An	nan vanished last Saturday.
	nceptual Structure:
	subconcept(weekday(Saturday):21,[time_period-1])
	subconcept(last:16,[last(a)-1,last-2,concluding-1,final-2,last-5,last-6,final-3,last-8,last-9])
	subconcept(date:weekday(Saturday):21,[Saturday-1])
	role(subsective,weekday(Saturday):21,last:16)
	role(degree,last:16,normal) subconcept(vanish:7(disappear-1)vanish-2,fly-8,vanish-4,vanish-5])
	role(Theme, vanish: 7, man:2)
	subconcept(man:3,[man-1,serviceman-1,man-3,homo-2,man-5,man-6,man-7,valet-1,Man-9,man-10,world-8]
	role(cardinality_restriction,man:3,sg)
	ntextual Structure:
	context(t)
	top_context(t)
	instantiable(date:weekday(Saturday):21,t)
	instantiable(man:3,t)
	instantiable(vanish:7,t)
	instantiable(weekday(Saturday):21,t)
Ten	nporal Structure:
	temporalRel(startsAfterEndingOf,Now,vanish:7)
	temporalRel(temporallySubsumes,date:weekday(Saturday):21,vanish:7)



00	🗙 parc bridge
	Display Analyses: 🔶 single 🔷 packed 🛛 Clear 🛛 Exit 🛛 Help
Passage	A man vanished last Saturday.
	Syntax AKR0 AKR
Question	Did a person disappear?
ducouon	
	Syntax AKR0 AKR
Answe	YES: [person=man,disappear=vanish]

000	X parc bridge
	Display Analyses: 🔶 single 🔷 packed 🛛 Clear 🛛 Exit 🛛 Help
Passage	The president believes that Iran has some uranium.
	Syntax AKR0 AKR
Question	Does Iran have uranium?
	Syntax AKR0 AKR
Answe	er UNKNOWN

000	🔀 parc bridge
	Display Analyses: 🔶 single 🔷 packed 🛛 Clear 🛛 Exit 🛛 Help
Passage	The president knows that Iran has some uranium.
	Syntax AKR0 AKR
Question	Does Iran have uranium?
	Syntax AKR0 AKR
Answe	r VES: [cordinality_comp(cg)]
Answe	r YES: [cardinality=some(sg)]

000	X parc bridge
	Display Analyses: 🔶 single 🔷 packed 🛛 Clear 🛛 Exit 🛛 Help
Passage	Sandy knows that Kim is not coming home.
	Syntax AKR0 AKR
Question	Is Kim coming home?
	Syntax AKR0 AKR
Answe	er NO

000	X parc bridge
	Display Analyses: 🔶 single 💊 packed 🛛 Clear 🛛 Exit 🛛 Help
Passage	The destruction of Carthage occurred in 149 BC.
	Syntax AKR0 AKR
Question	Was Carthage destroyed?
	Syntax AKR0 AKR
Answe	er YES

## Application: CALL

- The following demo is based on the large English LFG grammar.
- The grammar was modified
  - to parse ungrammatical sentences
  - register and communicate the type of ungrammaticality
  - generate the grammatical version

Khader, R. 2003. Evaluation of an English LFG-based Grammar as Error Checker. UMIST MSc Thesis, Manchester.







### Generation: CALL

- So far we have seen only the parsing capability of the grammar.
- But all LFG/XLE grammars can use the grammar to generate.
- For a given f-structure, the grammar goes through all the available rules to produce the possible output specified by the grammar.
- The LFG/XLE grammars can work with underspecified input this is what happens in the case with ungrammatical input.


# XLE in the Shell

This is what XLE looks like when called up from a command line:

% xle

XLE loaded from /usr/local/xle/bin/xle. XLEPATH = /usr/local/xle. Copyright (c) 1993-2001 by the Xerox Corporation and Copyright (c) 2002-2009 by the Palo Alto Research Center. All rights reserved. This software is made available AS IS, and PARC and the Xerox Corporation make no warranty about the software, its performance or its conformity to any specification. XLE version 2.6.4 (built Aug 10, 2009 09:39 -0700) Type 'help' for more information. loading /Users/mutt/pargram/call/english-call.lfg... Grammar has 382 rules with 8736 states, 22764 arcs, and 29349 disjuncts (36235 DNF).

## XLE in the Shell

This is what it looks like when the grammar generates from an f-structure.

% regenerate "They hope graduate in June." parsing {They hope graduate in June.} \*2+40 solutions, 0.050 CPU seconds, 0.000MB max mem, 535 subtrees unified

They hope to graduate in June.

regeneration took 0.12 CPU seconds. %

# Paradigms via Underspecified Generation

- Next is an example in which the grammar has been asked to generate without any tense/aspect information specified as part of the input.
- The result is the generation of the entire verbal paradigm of the language.

```
% regenerate "John sleeps."
parsing {John sleeps.}
1+5 solutions, 0.020 CPU seconds, 0.000MB max mem, 67 subtrees unified
```

### John

```
{ { will {have been|be}}
 |was
 |{has|had} been
 |is}
sleeping.
|will {have slept.|sleep.}
|{has|had} slept.
|slept.
|sleeps.}
```

```
regeneration took 0.09 CPU seconds. \%
```

# More CALL Potential

- The following demo is based on work by Melanie Seiss on the Australian Aboriginal language Murrinh-Patha.
- The system uses:
  - the large English LFG grammar
  - an LFG grammar and FSM for Murrinh-Patha
  - the XLE/XFR transfer system for machine translation
  - Perl scripts and a TCL/TK Interface

Seiss, Melanie & Rachel Nordlinger. 2011. An Electronic Dictionary and Translation System for Murrinh-Patha. *Proceedings of the EUROCALL* 2011 Conference, University of Nottingham.







	$\Delta$	
Which English sentence do you want to translate?		
the girls saw the boys		
Translate		
Is the subject a group of 2, a small group (ca. 3 - 10) or a bigger group (over ca. 10)?		
🔶 a group of two < a small group (ca. 3-10) < a big group (over ca 10)		
Are they siblings?		
♦ yes ◇ no What are siblings?		
Is the object a group of 2 people, a small group (ca. 3 - 10) of people, a bigger group (over ca. 10) of people or nonhuman?		
< a group of two 🔶 a small group < a big group < non-human		
Are they siblings?		
♦ yes ♦ no What are siblings?		
Are they male or female or mixed?		
💠 male 🛛 💠 female / mixed		
Clear Close Window Exit		

#### Translation Result:

English: Murrinh-Patha:	the girls saw the boys Kardu ngalarru pubampunkungkarduneme kardu kigay.				
Alternative:	Kardu ngalarru pubamkangkardu kardu kigay.				
Verb only: Pubampunkungkarduneme.					
Alternative:					



000	X More Information
	More information for 'kardu ngalarru pubampunkungkarduneme kardu kigay':
	Get dictionary entry and examples
	Get morphological information
	Show form with all possible tenses
	Show form with all possible subject numbers
	Show form with all possible object numbers
	Close Window

000	More Information
	More information for 'kardu ngalarru pubampunkungkarduneme kardu kigay':
	Get dictionary entry and examples
	Get morphological information
Morpholo	gy for kardu ngalarru pubampunkungkarduneme kardu kigay:
kardu	Noun Classifier
ngalarru	Noun
pubam punku ngkardu neme pldunFut: kardu kigay	Classifier 13, Subject Information: 3. Person dual Sibling (they two Siblings), Tense: non-Future Direct Object Information: 3 Person dual Sibling (them, two siblings) Lexical stem Object Information: few male non-Sibling (overwrites info) Subject: few female non-Siblings (overwrites info from classifier) Subject: few male non-Siblings (overwrites info from classifier) Noun Classifier
	Show form with all possible tenses Show form with all possible subject numbers Show form with all possible object numbers

### Show form with all possible tenses

Non-Future: Past Imperfective: Future: Futur Irrealis: Past Irrealis: Kardu ngalarru pubampunkungkarduneme kar Kardu ngalarru pubenkungkardudhaneme karc Kardu ngalarru pubankungkardununeme kardu Kardu ngalarru kubankungkardunukunneme karc

# Deep Grammars

- Deep Grammars have great potential with respect to NLP applications.
- Deep Grammars serve as a systematic tool for language documentation
- Why don't more people use them?

# Disadvantages

- Time consuming and expensive to write
  - shallow parsers can be induced automatically from a training set (but somebody has to construct that first)
- Brittle
  - shallow parsers produce something for everything
- Ambiguous
  - shallow parsers rank the outputs
- Slow
  - shallow parsers are very fast (real time)

# Why pay attention now?

### **New Generation of Large-Scale Grammars:**

- Robustness:
  - Integrated Chunk Parsers/Fragment Grammars
  - Bad input always results in some (possibly good) output
- Ambiguity:
  - Integration of stochastic methods
  - Optimality Theory used to rank/pick alternatives
- Speed: getting closer to shallow parsers
- Accuracy and information content:
  - far beyond the capabilities of shallow parsers.

## **XLE at PARC**

- Platform for Developing Large-Scale LFG Grammars
- LFG (Lexical-Functional Grammar)
  - Invented in the 1980s
     (Joan Bresnan and Ronald Kaplan)
  - Theoretically stable ⇔ Solid Implementation
- XLE is implemented in C, used with emacs, tcl/tk
- XLE includes a parser, generator and transfer (XFR) component.

## **XLE at PARC**

- Platform for Developing Large-Scale LFG Grammars
- LFG (Lexical-Functional Grammar)
  - Invented in the 1980s
     (Joan Bresnan and Ronald Kaplan)
  - Theoretically stable ⇔ Solid Implementation
- XLE is implemented in C, used with emacs, tcl/tk
- XLE includes a parser, generator and transfer (XFR) component.
- Extensive on-line documentation

### **Background Literature**

- Bresnan, Joan & Ron Kaplan. 1982. The Mental Representation of Grammatical Relations. MIT Press.
- Bresnan, Joan. 2001. Lexical-Functional Syntax. Blackwell.
- Dalrymple, Mary. 2001. Lexical-Functional Grammar. Academic Press.
- Butt, Miriam, Tracy Holloway King, Frédérique Segond & María-Eugenia Niño. 1999. *The Grammar Writer's Cookbook*. CSLI Publications.

### **XLE/LFG and ParGram**

- XLE has been used to develop a number of grammars for very different languages as part of the **ParGram** (Parallel Grammar) collaborative effort.
- (see websites, ParGram Wiki for detailed information)

## **The Parallel in ParGram**

- Analyze languages to a degree of abstraction that reflects the common underlying structure (i.e., identify the subject, the object, the tense, mood, etc.).
- Even at this level, there is usually more than one way to analyze a construction.
  - The same theoretical analysis may have different possible implementations.
  - ParGram agrees on common analyses and implementations (via regular meetings and a feature committee).

### The Parallel in ParGram

- Analyses at the level of c-structure are allowed to differ (variance across languages)
- Analyses at f-structure are held as parallel as possible across languages (crosslinguistic invariance).
- Theoretical Advantage: This models the idea of universal principles underlying language structure.
- Applicational Advantage: machine translation is made easier; applications are more easily adapted to new languages.

## ParGram Languages (so far)

- Arabic, Chinese, Danish, English, French, Georgian, German, Hungarian, Irish Gaelic, Indonesian, Japanese, Malagasy, Murrihn-Patha, Norwegian, Polish, Setswana, Tigrinya, Turkish, Urdu, Welsh, Wolof
- Loose organization: no common deliverables, but common interests.
- Recent Multilingual Effort: the ParGramBank
- Many grammars are available via the INESS website.

Sebastian Sulger, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Agnieszka Patejuk, Özlem Çetinoğlu, I Wayan Arka and Meladel Mistica: ParGramBank: The ParGram Parallel Treebank. *Proceedings of ACL 2013* (Long Papers), Sofia, Bulgaria. Association for Computational Linguistics.

# Getting Started with XLE

- This concludes today's lesson.
- As the practical part of this lesson, you should now:
  - I. Install XLE so that you can use it for the exercises in the course.
  - 2. Go to the INESS website where many of the ParGram grammars are available and experiment with the grammars and the representations.
  - 3. The English LFG grammar is particularly large and robust.
- Detailed information and all the necessary links are provided separately in the exercise file for the lesson.

