Machine Translation
ICS 482 Natural Language Processing

Lecture 29-2: Machine Translation

Husni Al-Muhtaseb

بسم الله الرحمن الرحيم

ICS 482 Natural Language Processing

Lecture 29-2: Machine Translation

Husni Al-Muhtaseb

NLP Credits and Acknowledgment

These slides were adapted from presentations of the Authors of the book

SPEECH and LANGUAGE PROCESSING:

An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition

and some modifications from presentations found in the WEB by several scholars including the following

NLP Credits and Acknowledgment

If your name is missing please contact me
muhtaseb
At
Kfupm.
Edu.
sa

NLP Credits and Acknowledgment

Husni Al-Muhtaseb
James Martin
Jim Martin
Dan Jurafsky
Sandiway Fong
Song young in
Paula Matuszek
Mary-Angela Papalaskari
Dick Crouch
Tracy Kin
L. Venkata Subramaniam
Martin Volk
Bruce R. Maxim
Jan Hajič
Srinath Srinivasa
Simeon Ntafos
Paolo Pirjanian
Ricardo Vilalta
Tom Lenaerts

Heshaam Feili
Björn Gambäck
Christian Korthals
Thomas G. Dietterich
Devika Subramanian
Duminda Wijesekera
Lee McCluskey
David J. Kriegman
Kathleen McKeown
Michael J. Ciaraldi
David Finkel
Min-Yen Kan
Andreas Geyer-Schulz
Franz J. Kurfess
Tim Finin
Nadjet Bouayad
Kathy McCoy
Hans Uszkoreit
Azadeh Maghsoodi

Khurshid Ahmad
Staffan Larsson
Robert Wilensky
Feiyu Xu
Jakub Piskorski
Rohini Srihari
Mark Sanderson
Andrew Elks
Marc Davis
Ray Larson
Jimmy Lin
Marti Hearst
Andrew McCallum
Nick Kushmerick
Mark Craven
Chia-Hui Chang
Diana Maynard
James Allan

Martha Palmer
julia hirschberg
Elaine Rich
Christof Monz
Bonnie J. Dorr
Nizar Habash
Massimo Poesio
David Goss-Grubbs
Thomas K Harris
John Hutchins
Alexandros
Potamianos
Mike Rosner
Latifa Al-Sulaiti
Giorgio Satta
Jerry R. Hobbs
Christopher Manning
Hinrich Schütze
Alexander Gelbukh
Gina-Anne Levow
Guitao Gao
Qing Ma
Zeynep Altan

Today's Lecture

- Machine Translation (MT)
  - Structure of Machine Translation System
  - A simple English to Arabic Machine Translation

Structure of MT Systems

- Generally they all have lexical, morphological, syntactic and semantic components, one for each of the two languages, for treating basic words, complex words, sentences and meanings

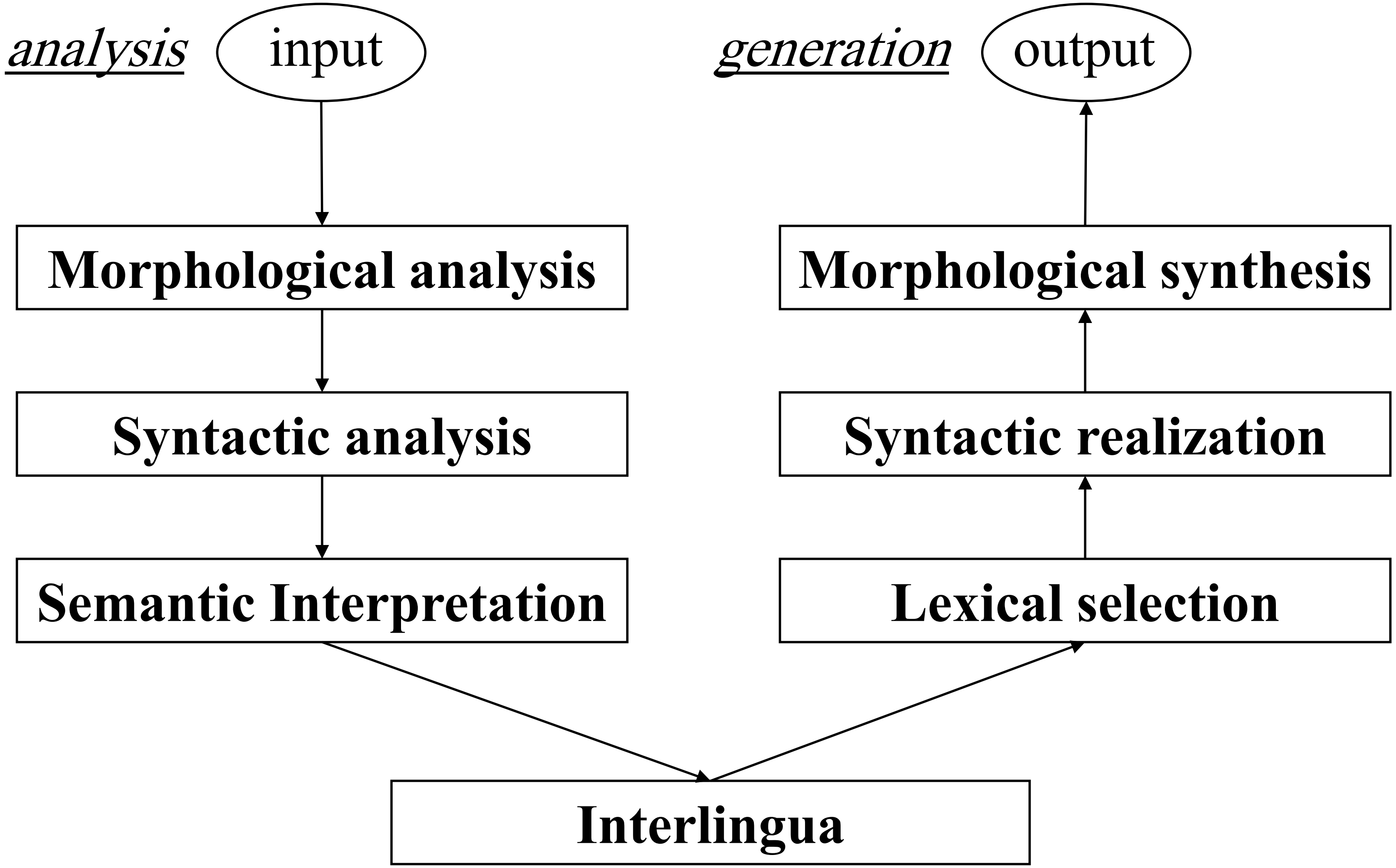
Structure of MT Systems(cont.)

- “transfer” component: the only one that is specialized for a particular pair of languages, which converts the most abstract source representation that can be achieved into a corresponding abstract target representation

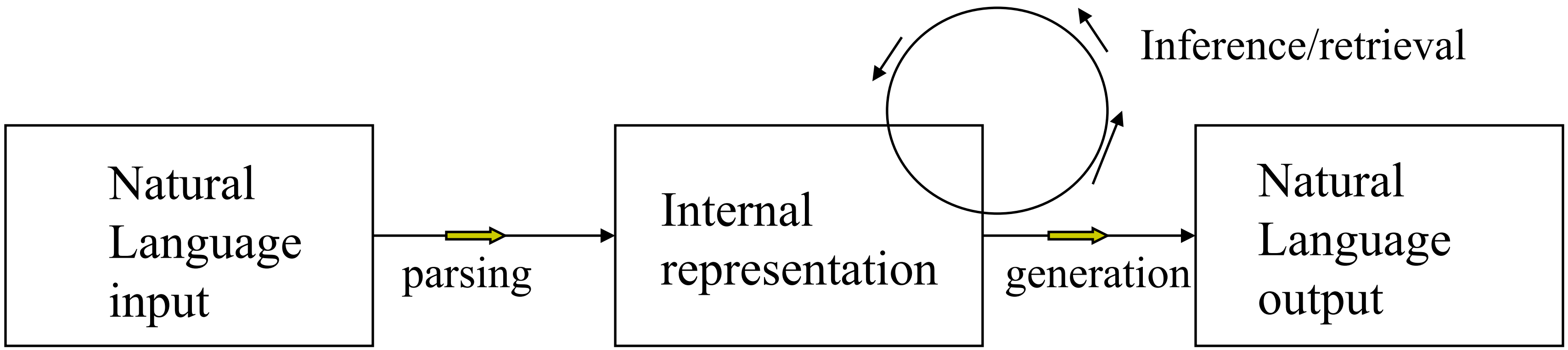
Structure of MT Systems(cont.)

- Some systems make use of a so-called "interlingua" or intermediate language
  - The transfer stage is divided into two steps, one translating a source sentence into the interlingua and the other translating the result of this into an abstract representation in the target language

Machine Translation

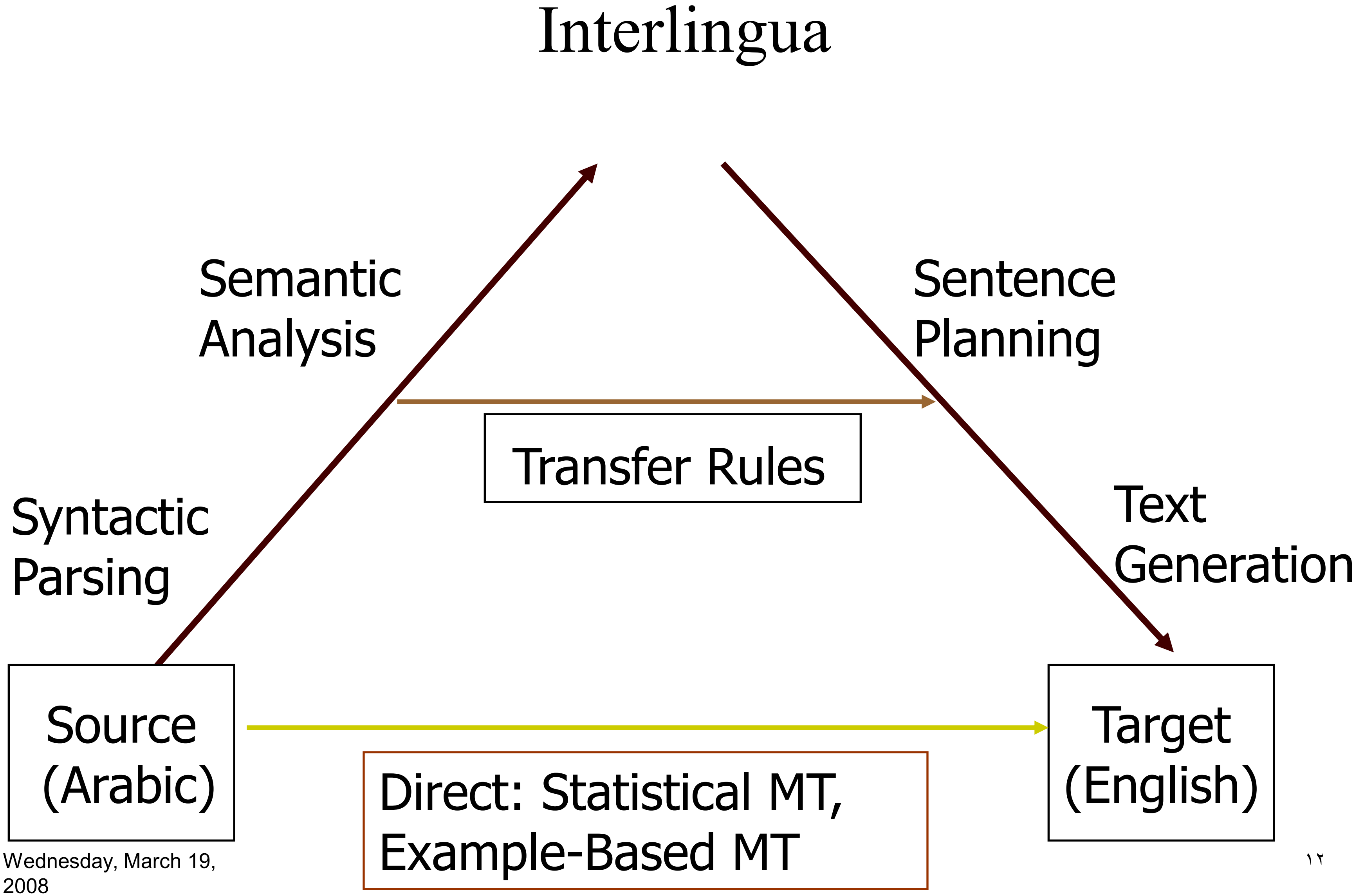


Typical NLP System

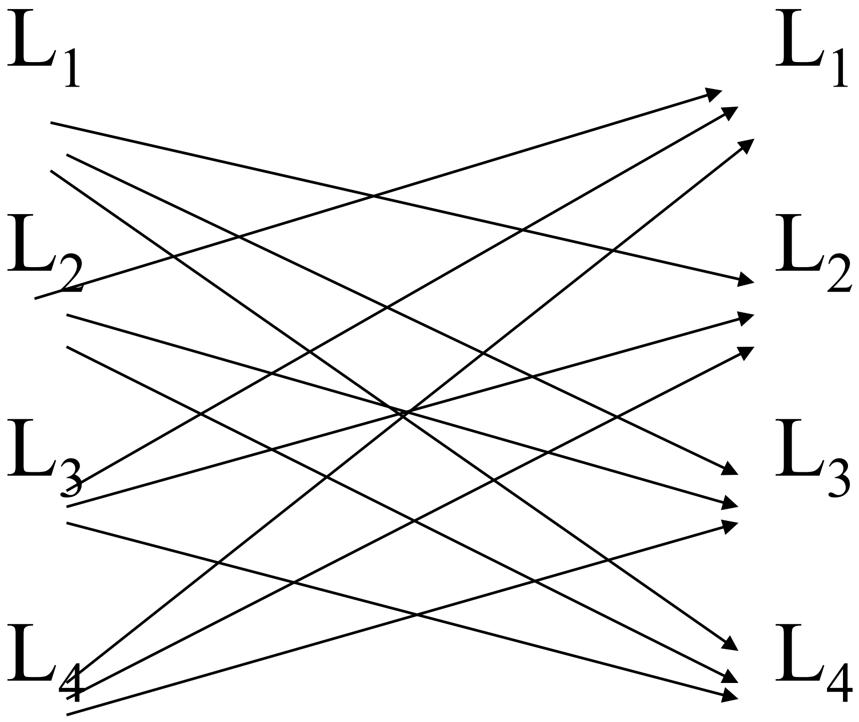


- NL Data-Base Query:
  - Parsing = Question → SQL query
  - Inference/retrieval = DBMS: SQL → table of records
  - Generation = no-operation (just print the retrieved records)
- Machine Translation
  - Parsing = Source Language text → Representation
  - Inference/retrieval = no-operation
  - Generation = Representation → Target language

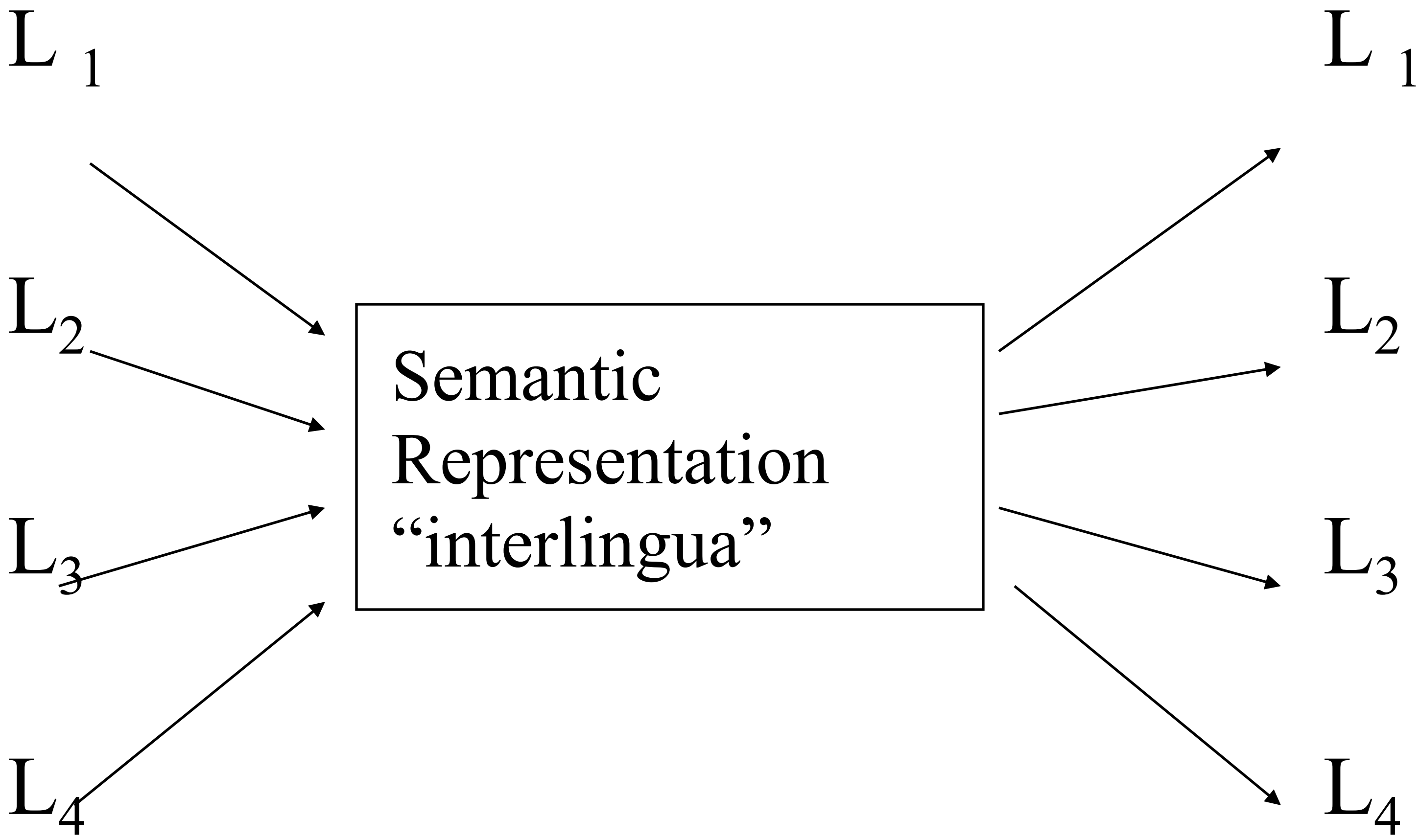
Types of Machine Translation



Transfer Grammars



Interlingua Paradigm for MT



Interlingua-Based MT

- Requires an Interlingua - language-neutral Knowledge Representation (KR)
  - Philosophical debate: Is there an interlingua?
  - FOL is not totally language neutral (predicates, functions, expressed in a language)
  - Other near-interlinguas (Conceptual Dependency)
- Requires a fully-disambiguating parser
  - Domain model of legal objects, actions, relations
- Requires a NL generator (KR → text)
- Applicable only to well-defined technical domains
- Produces high-quality MT in those domains

Example-Based MT (EMBT)

- Can we use previously translated text to learn how to translate new texts?
  - Yes! But, it's not so easy
  - Two paradigms, statistical MT, and EBMT
- Requirements:
  - Aligned large parallel corpus of translated sentences

    $\{S_{source} \Leftrightarrow S_{target}\}$
  - Bilingual dictionary for intra-S alignment
  - Generalization patterns (names, numbers, dates…)

EBMT Approaches

- Simplest: Translation Memory
  - If $S_{new} = S_{source}$ in corpus, output aligned S_{target}
- Compositional EBMT
  - If fragment of S_{new} matches fragment of S_s, output corresponding fragment of aligned S_t
  - Prefer maximal-length fragments
  - Maximize grammatical compositionality
    - Via a target language grammar,
    - Or, via an N-gram statistical language model

Multi-Engine Machine Translation

- MT Systems have different strengths
  - Rapidly adaptable: Statistical, example-based
  - Good grammar: Rule-Based (linguistic) MT
  - High precision in narrow domains: INTERLINGUA
- Combine results of parallel-invoked MT
  - Select best of multiple translations

Our Approach: Structure of Translator

- Lexical Module
- Syntax Module
- Transformation Module

Lexical Module

- Pre Processor
  - Detect Proper Nouns
  - Convert short forms (don’t → do not)
  - Detect abbreviations like etc., mr.
- Tokenizer

  Search Database of words and proper nouns and generate all possible interpretations of a word.

Structure of Lexicon

- Word
- Category
  - Noun, Pronoun, …
- Subcategory
  - Auxiliary Verb, Possessive Pronoun, ToPreposition, …
- Sense
  - Human, Animate, Unanimate

Structure of Lexicon - *Contd.*

- Form
  - Base, First,Second, … (for Verb Form); First, Second,Third (for Person); Comparative, Superlative, … for Adjectives
- Number
  - Singular, Plural
- Gender
  - Masculine, Feminine
- Object Preposition & Subject Preposition

Structure of Lexicon - *Contd.*

- Object Count
  - Number of objects required with the verb
- Arabic Meaning
- Meaning for different forms
  - Meaning of Adjective and Noun for different forms of Gender and Number

English to Arabic Machine Translation

- Salma came
- Lexicon
  - Salma:.. سلمى، اسم علم، مؤنث، مفرد،
  - Came: ... جاء، فعل، ماض، متعادل
- Word to word: سلمى جاء
- Needed Translation: جاءت سلمى
- Modification Rules
  - Exchange the positions of subject and verb
  - If the gender is feminine the verb should be the same

A second Example

- The students are active
- Lexicon
  - The: ال
  - Students: طلاب، اسم جنس، جمع، متعادل، ..
  - Are: فعل، مضارع، يكون، جمع، متعادل ..
  - Active: صفة، نشيط، متعادل، ..
- Word to Word: ال طلاب يكون نشيط
- Needed Translation: الطلاب نشيطون
- Modification Rules
  - Insert ال with its successor
  - Omit يكون
  - Change نشيط to proper number (plural) and proper gender (masculine)
- What about: Needed Translation: الطالبات نشيطات

More Examples

- Lena had recently added a home-theater sound system to the TV
  - لينا قام مؤخرا اضاف منزل-مسرح صوت نظام الى التلفاز
  - قامت لينا مؤخرا بإضافة نظام صوت مسرح-منزلي الى التلفاز.
- The fans in the stand were screaming
  - ال مشجعون في ال منصة كانوا صراخ
  - المشجعون في المنصة كانوا يصرخون.
  - كان المشجعون في المنصة يصرخون.

Final Exam - Related

- NLP Repeated Concepts
  - Things you should know by now
- Lectures 12 – Today's Lecture
  - Related Material from the book
    - From Chapters 10, 12, 14, 15, 16, 21
- Take Home Quiz & Related Material
- Student Presentations
  - Main Concepts
  - Student Questions
  - Your presentation
- Your team project
- No Final Exam Sample

Thank you

أسأل الله أن يعيننا وإياكم وأن يوفق الجميع إلى كل خير

سبحانك اللهم وبحمدك، أشهد أن لا إله إلا أنت، أستغفرك وأتوب إليك

السلام عليكم ورحمة الله

Thank you

- السلام عليكم ورحمة الله