Natural Language Processing

What Is NLP ?

- **“Natural” languages**
  - English, Mandarin, French, Swahili, Arabic, Nahuatl, ...
  - NOT Java, C++, Perl, ...

- **Natural language processing (NLP)**
  - **WIKI**: NLP is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (**natural**) languages.

- **NLP = NLU + NLG**
  - NLU = Natural Language Understanding, speech/text → meaning
  - NLG = Natural Language Generation, meaning → text/speech

Course Goals

- Introduction to the know-how of NLP (NLU), including
  - research highlights,
  - crucial technologies and
  - application achievements;
- Providing a chance to train students for reading and evaluating new academic
- Encouraging students to present and discuss their comments for the papers.
- Accomplish a practical NLP system through a course project.

Knowledge Requirement for Machine

- Phonetics and Phonology: knowledge about linguistic sounds
- Morphology: knowledge of the meaningful components of words
- Syntax: knowledge of the structural relationships between words
- Semantics: knowledge of meaning
- Pragmatics: knowledge of the relationship of meaning to the goals and intentions of the speaker
- Discourse: knowledge about linguistic units larger than a single utterance

Phonetics and Phonology

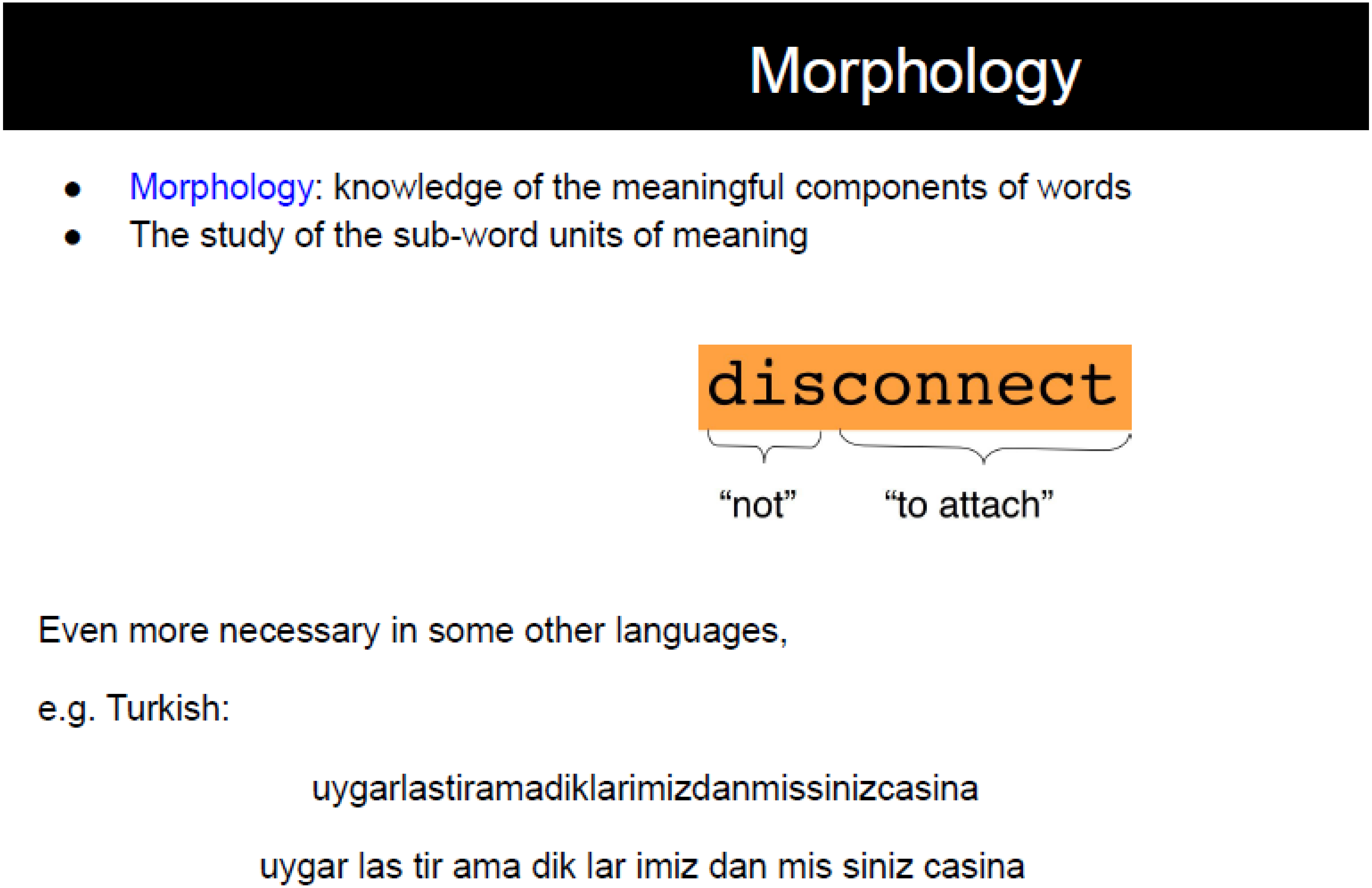
- Phonetics and Phonology: knowledge about linguistic sounds
- The study of:

  language sounds how they are physically formed;

  systems of discrete sounds, e.g. languages′ syllable structure

dis-k&-'nekt disconnect

Morphology التصريف او الأشتقاق



Morphology

- Morphology: knowledge of the meaningful components of words
- The study of the sub-word units of meaning

disconnect

“not” “to attach”

Even more necessary in some other languages,

e.g. Turkish:

uygarlastiramadiklarimizdanmissinizcasina

uygar las tir ama dik lar imiz dan mis siniz casina

What is Morphology
التصريف

The study of how words are formed

Words are made up of smaller meaning-bearing units (morphemes).

Morpheme: smallest meaningful unit of a word. Morphemes can be: stems, suffixes, prefixes …)

المدرسين

سيكتبها

Books

Unliked

Morphology is the identification, analysis, and description of the morphemes of a given word, including its root, part-of-speech, gender, number, etc.

Types of Morphology : Inflectional Morphology and derivational morphology

Root and Stem

Root (جذر): is the primary lexical unit of word, and which **carries aspects of semantic content**.

- Roots in Arabic are **decided by linguists** (who sometimes disagree).
- Some words may have more than one root, سنة (س ن ي)(س ن و)

Stem (ساق): is the part that is common to all its inflected variants (no agreed definition), thus it **depends on the stemming algorithm**.

Roots and Stems in English are most likely the same, but in Arabic Roots and Stems are not the same.

Examples:

Word	Lemma	Root	Stem
سيكتبها	كَتَبَ	ك ت ب	كتب
مكتباتهم	مَكْتَبَةٌ	ك ت ب	مكتب
مكاتبكم	مَكْتَبٌ	ك ت ب	مكاتب
كتابي	كِتَابٌ	ك ت ب	كتاب

English Derivational Morphology (الاشتقاق بالإنجليزية)

Examples of derivational patterns in English

verb-to-noun (agent):	er	write → writer
verb-to-adjective:	able	eat → eatable
verb-to-noun:	ance	guide → guidance
noun-to-adjective:	al	relation → relational
noun-to-verb:	fy	test → testify
adjective-to-noun:	ness	weak → weakness
adjective-to-verb:	en	weak → weaken
adjective-to-adjective:	ish	red → reddish
adjective-to-adverb:	ly	relation → relationally
….	…	….

Part-of-Speech (اقسام الكلام)

Also called **word class**, **lexical class**, and **lexical category**

a category of words that have similar grammatical properties.

Words have the same POS have similar syntaxic behavior, and sometimes similar morphology

In English:

- **noun** (e.g., book)
- **verb** (e.g., booked)
- **adjective** (e.g.,)
- **adverb** (e.g., very, quite)
- **pronoun** (e.g., he, them)
- **preposition** (e.g., in, of)
- **conjunction** (e.g., and, but)
- **interjection** (e.g., ops, alas)
- **numeral** (e.g., one, two)
- **article** (e.g., the, a , an)

In Arabic:

- **noun**
- **verb (PV, IV, CV)**
- **Letter/Functional words**

  words that are not nouns or verbs)

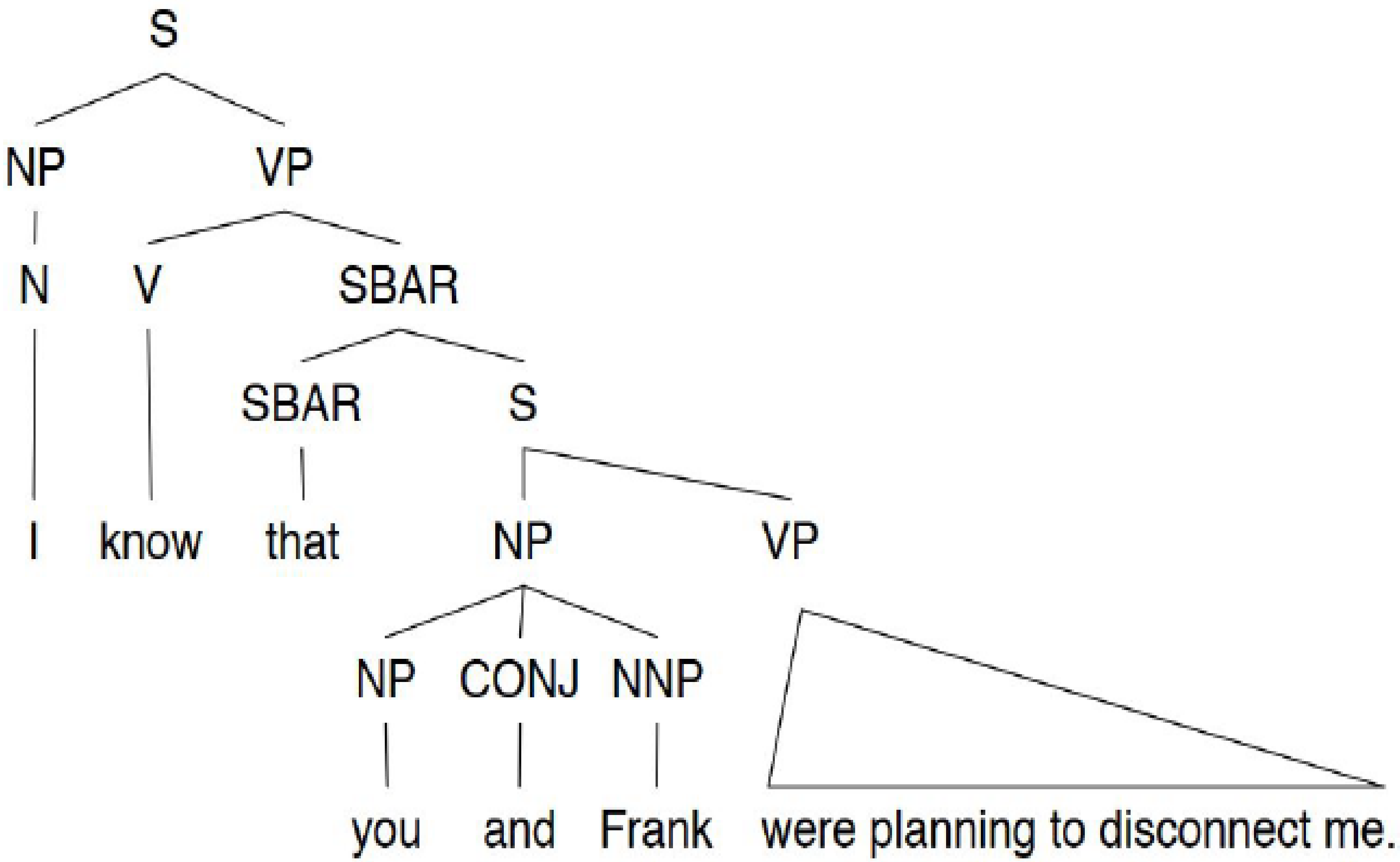
Next Lecture

➔Part-of-Speech Tagging

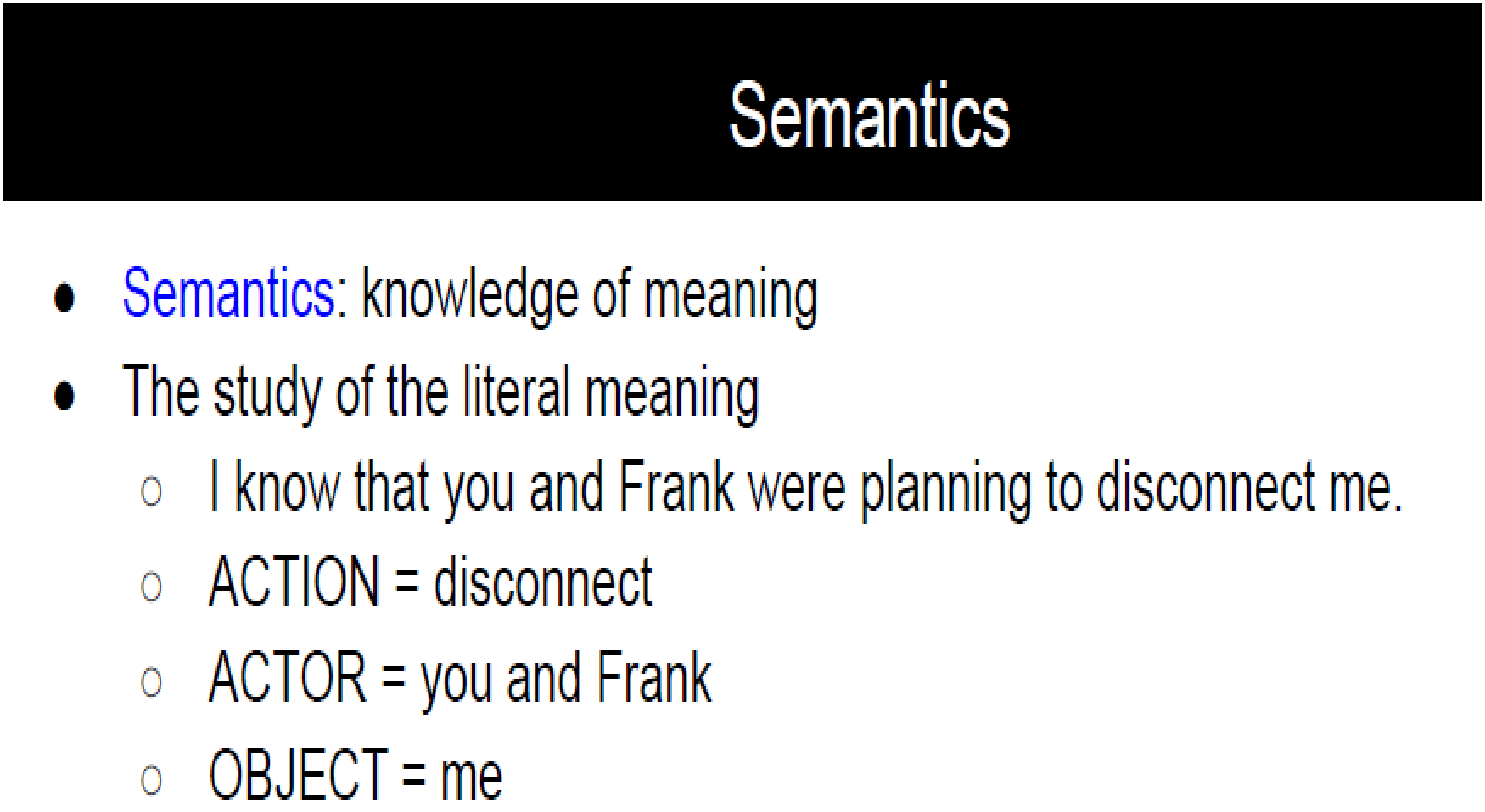
Syntax تراكيب الجملة

Syntax

- Syntax: knowledge of the structural relationships between words
- The study of the structural relationships between words
  - I know that you and Frank were planning to disconnect me.



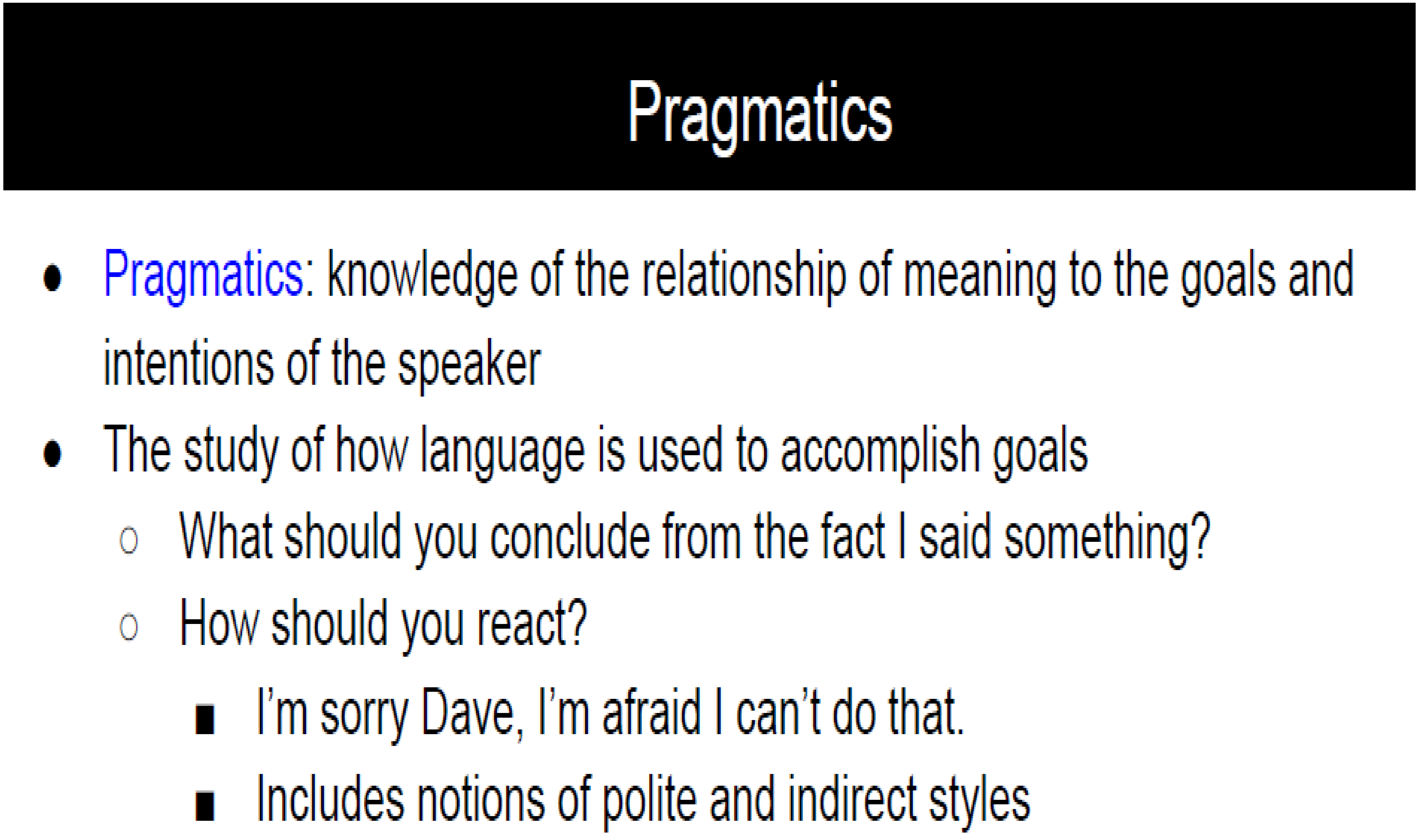
Semantics علم دلالات الألفاظ



Semantics

- Semantics: knowledge of meaning
- The study of the literal meaning
  - I know that you and Frank were planning to disconnect me.
  - ACTION = disconnect
  - ACTOR = you and Frank
  - OBJECT = me

في علم اللغويات، تُعنى البراغماتية بدراسة كيفية تأثير السياق على معنى اللغة، وكيفية استخدام الناس للغة في المواقف الاجتماعية للتواصل بما يتجاوز حدود الكلمات الحرفية.



Pragmatics

- Pragmatics: knowledge of the relationship of meaning to the goals and intentions of the speaker
- The study of how language is used to accomplish goals
  - What should you conclude from the fact I said something?
  - How should you react?
    - I'm sorry Dave, I'm afraid I can't do that.
    - Includes notions of polite and indirect styles

Semantics vs. Pragmatics

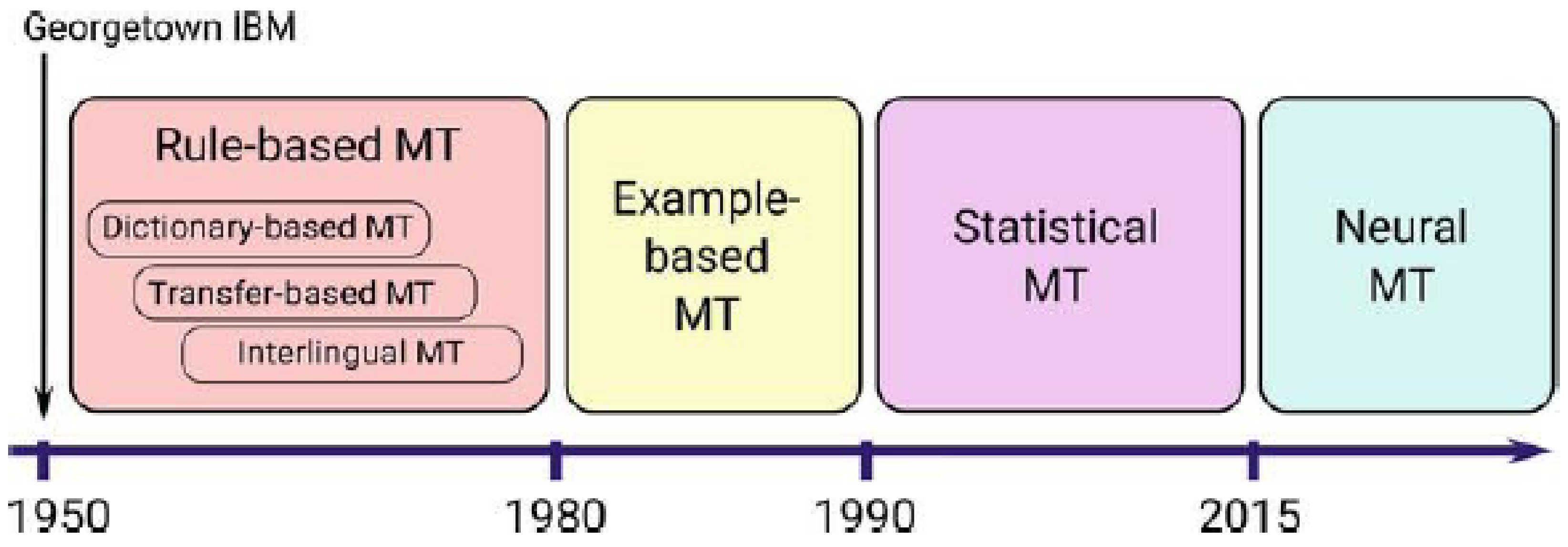
What does "You have a green light” mean?

- □ You are holding a green light bulb?
- □ You have a green light to cross the street?
- □ You can go ahead with your plan?

Tasks/Applications in NLP

Spoken Dialog Systems

Machine Translation



Georgetown IBM

Rule-based MT
- Dictionary-based MT
- Transfer-based MT
- Interlingual MT

Example-based MT

Statistical MT

Neural MT

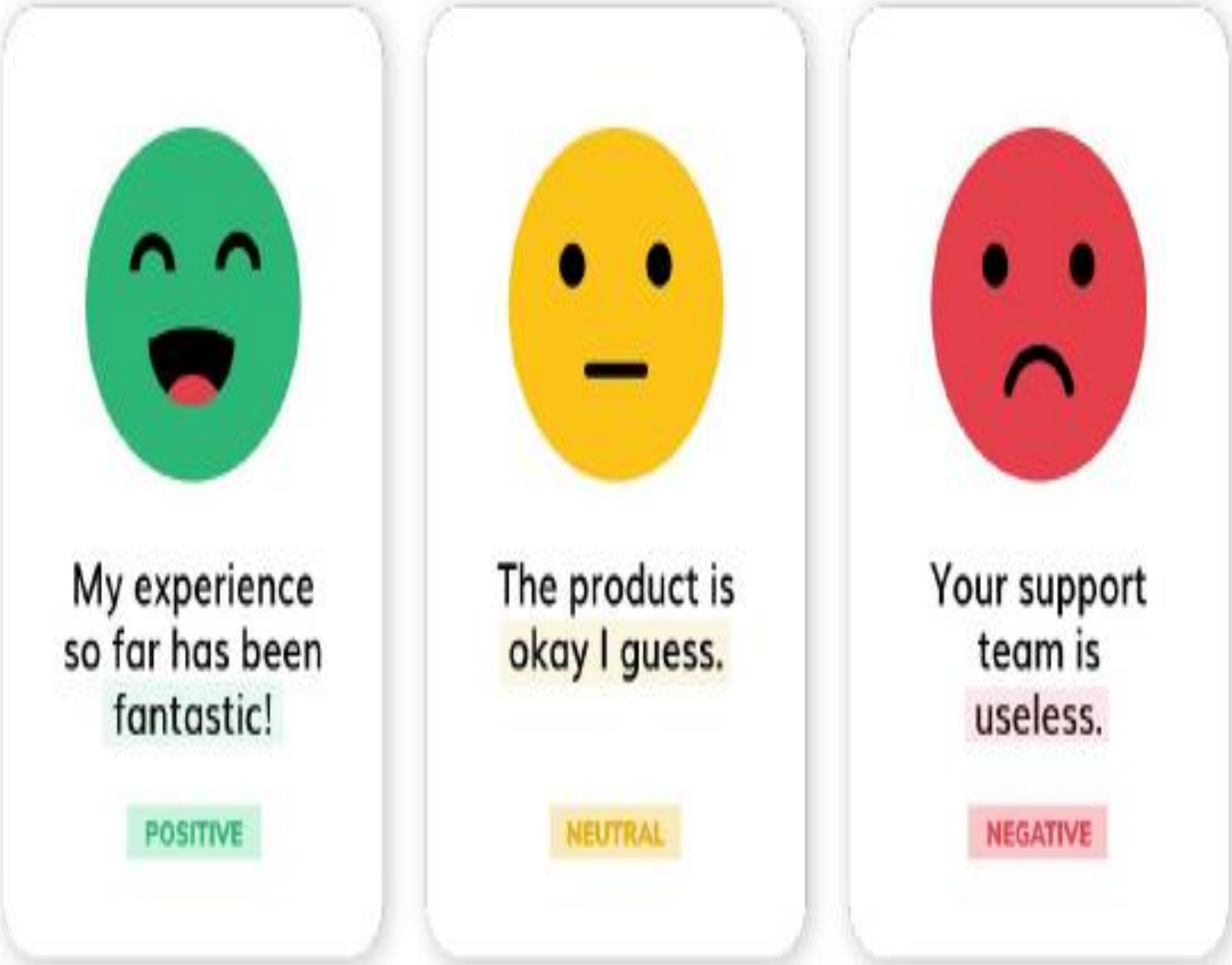
1950 1980 1990 2015

Low resource languages can be challenging?

6,800 living languages
600 with written tradition
100 spoken by 95% of population

Sentiment Analysis

Determine whether the meaning behind data is positive, negative, or neutral

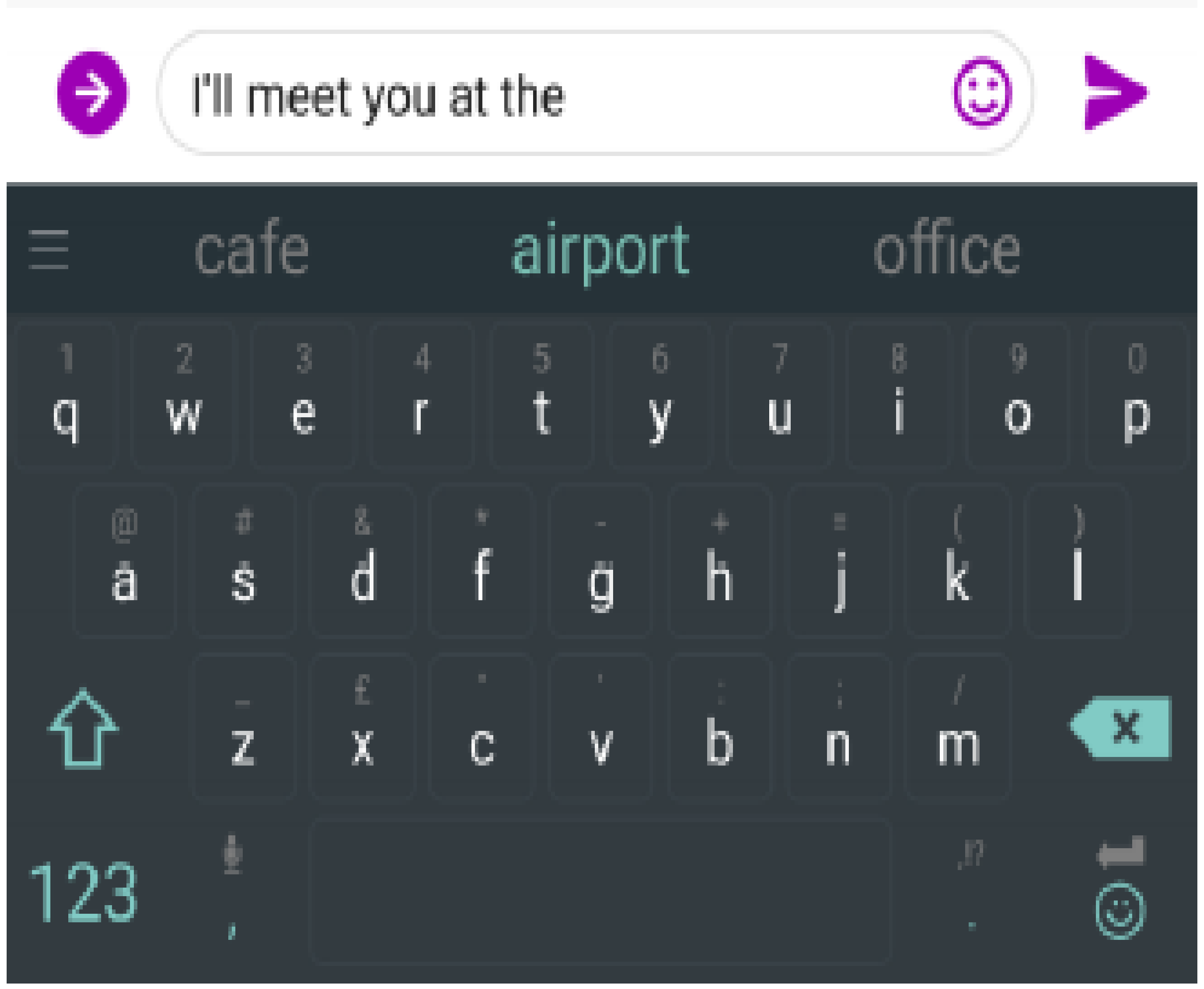


My experience so far has been fantastic!

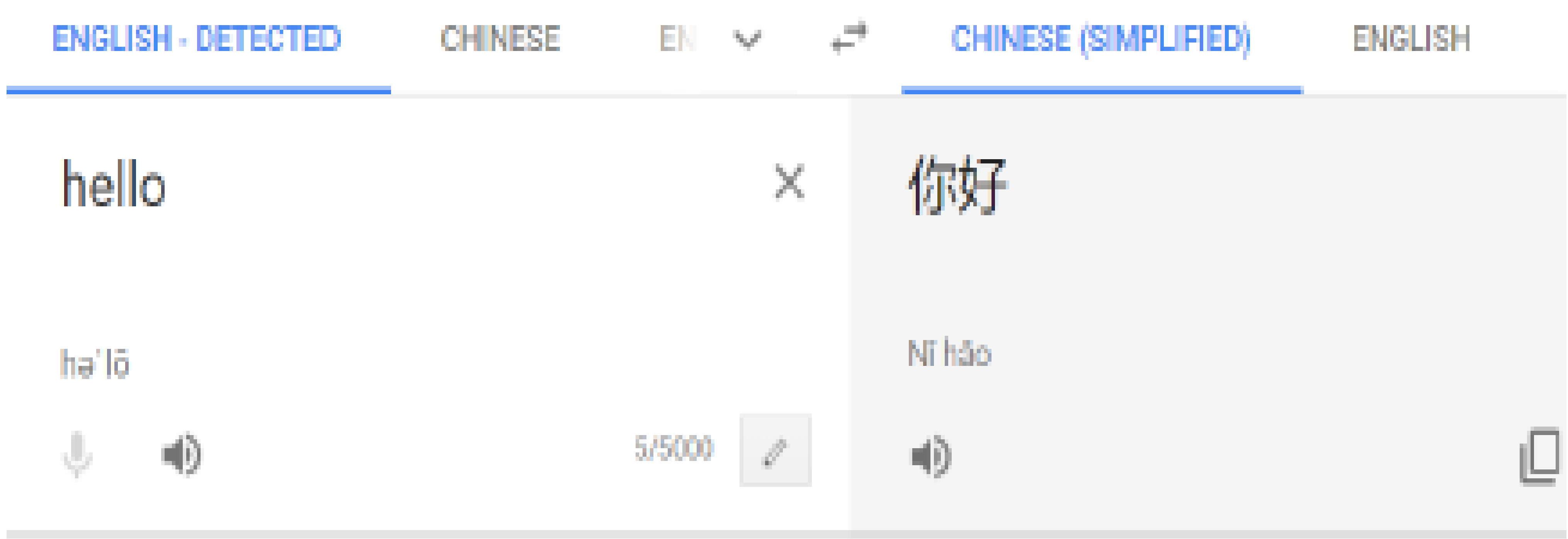
The product is okay I guess.

Your support team is useless.

Real-word NLP



I'll meet you at the

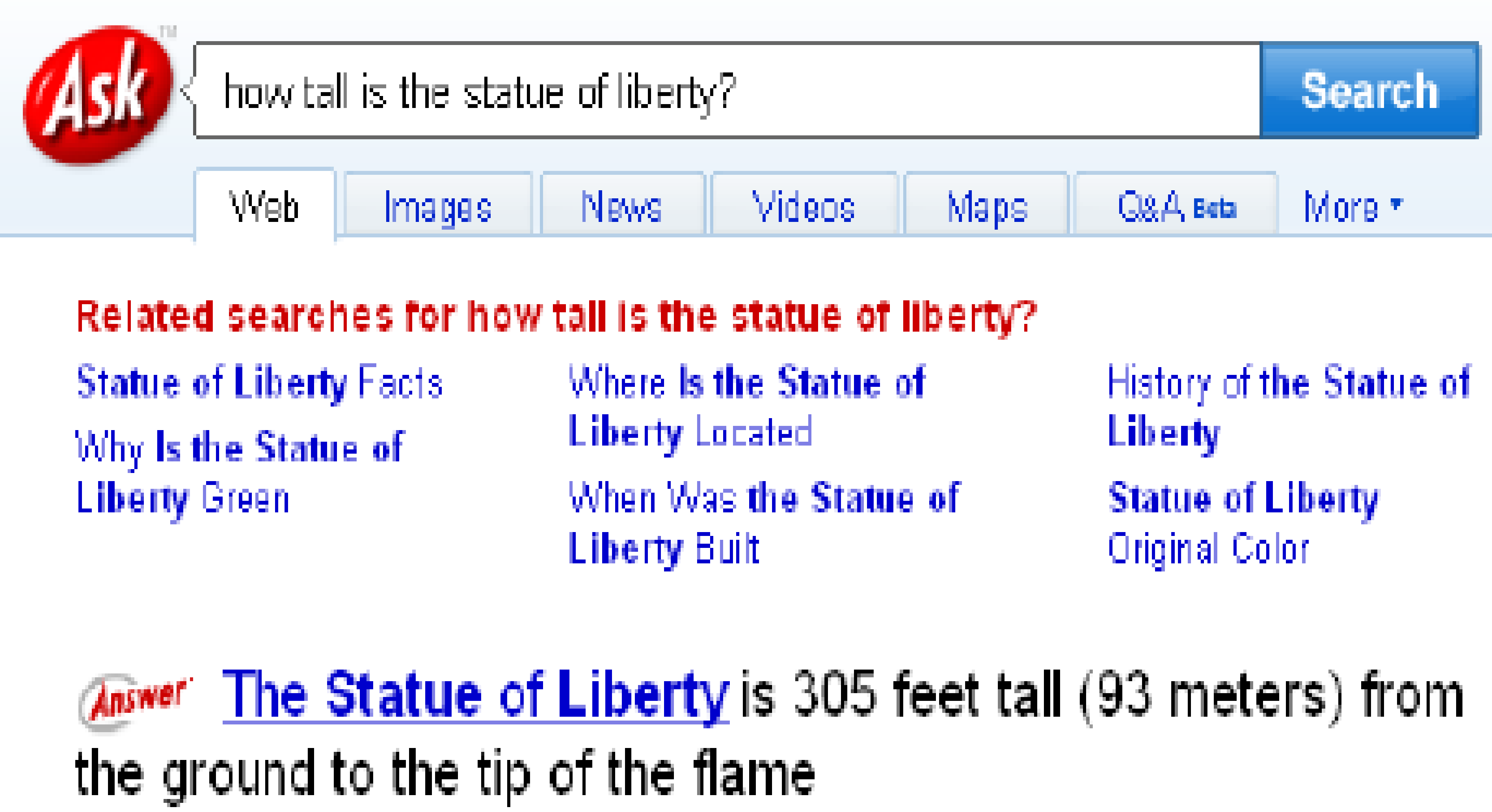


hello

hə'lō

你好

Nǐ hǎo



how tall is the statue of liberty?

Related searches for how tall is the statue of liberty?

Statue of Liberty Facts

Why Is the Statue of Liberty Green

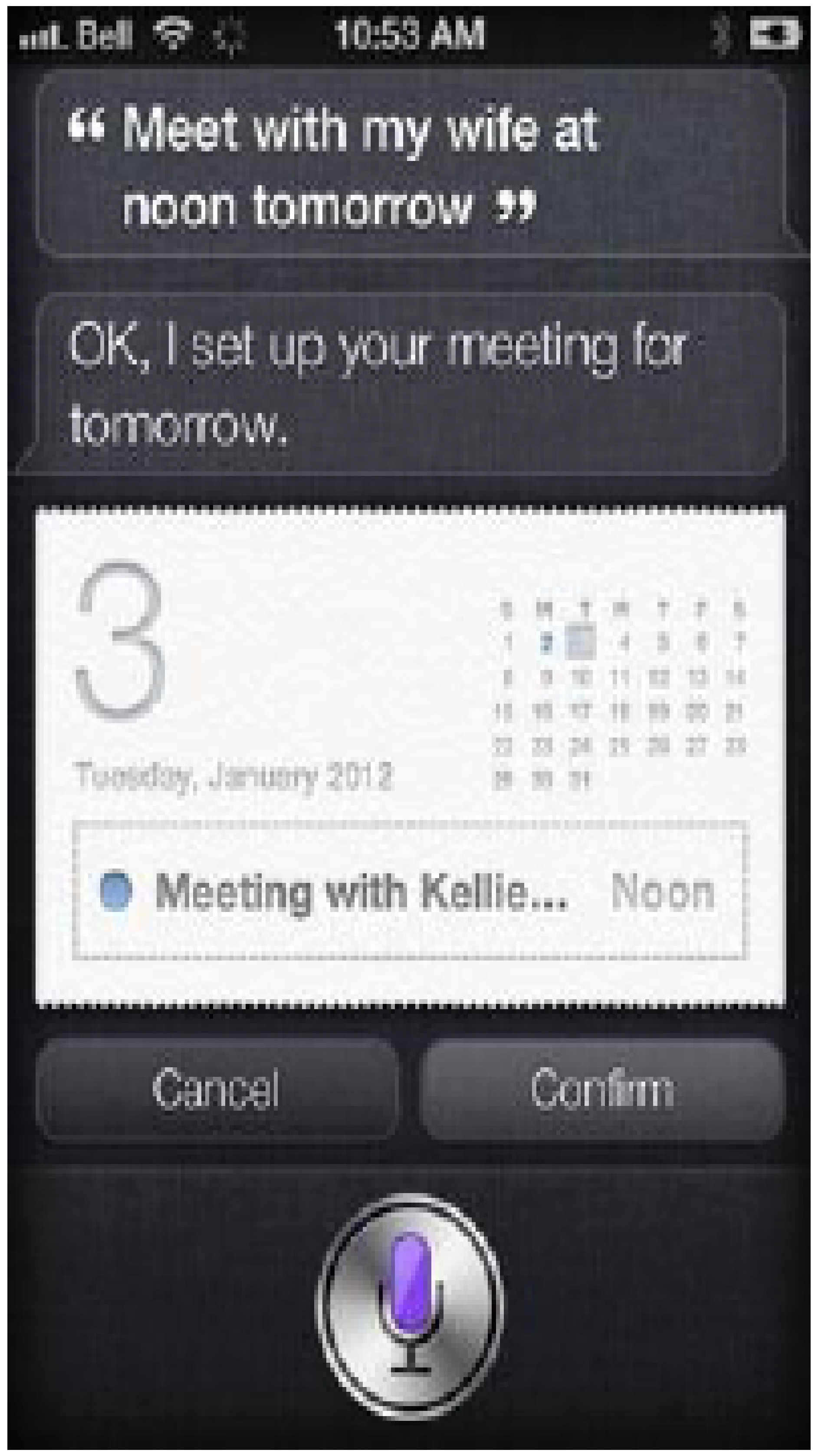
Where Is the Statue of Liberty Located

When Was the Statue of Liberty Built

History of the Statue of Liberty

Statue of Liberty Original Color

Answer The Statue of Liberty is 305 feet tall (93 meters) from the ground to the tip of the flame



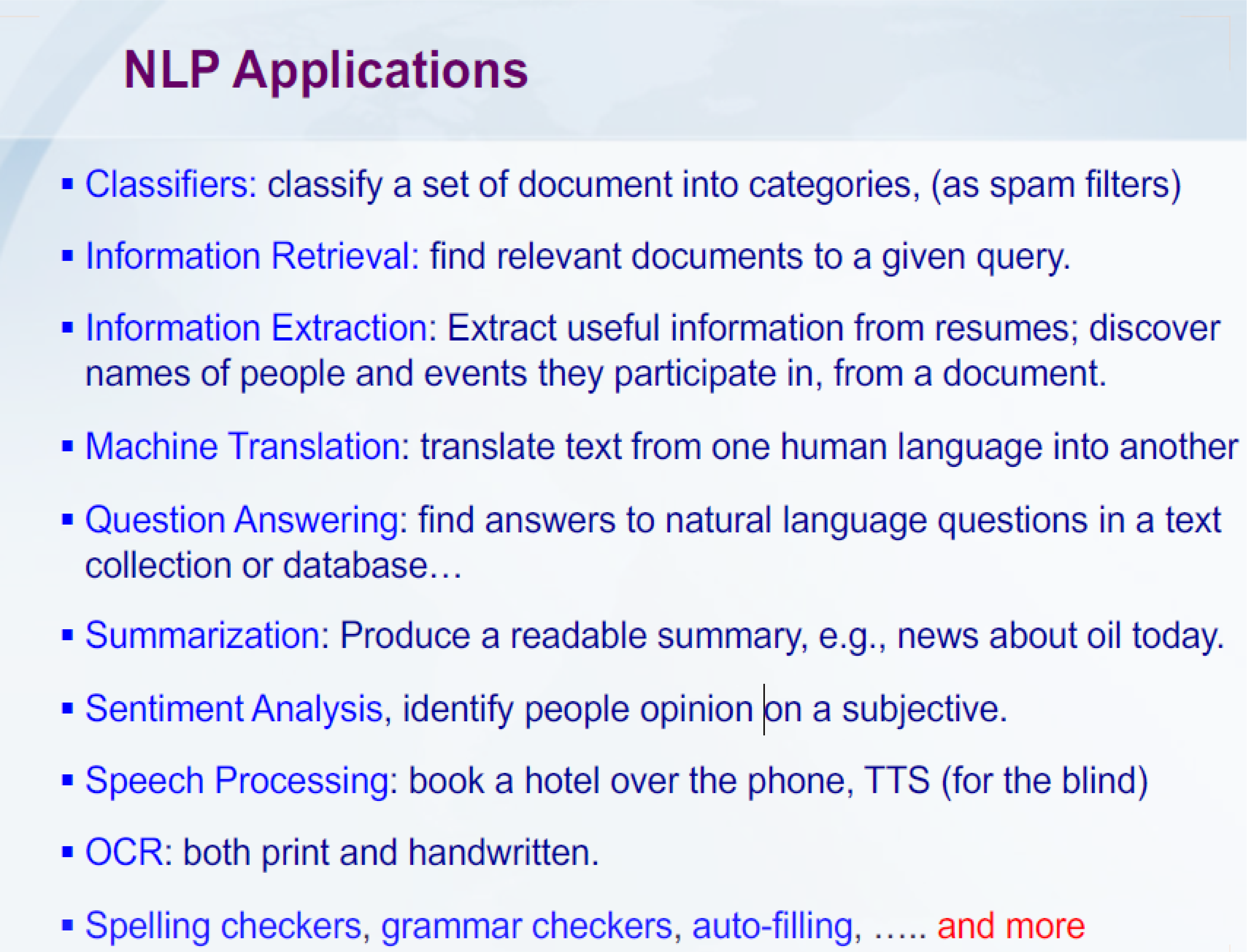
“Meet with my wife at noon tomorrow”

OK, I set up your meeting for tomorrow.

3

Tuesday, January 2012

Meeting with Kellie... Noon



NLP Applications

- Classifiers: classify a set of document into categories, (as spam filters)
- Information Retrieval: find relevant documents to a given query.
- Information Extraction: Extract useful information from resumes; discover names of people and events they participate in, from a document.
- Machine Translation: translate text from one human language into another
- Question Answering: find answers to natural language questions in a text collection or database…
- Summarization: Produce a readable summary, e.g., news about oil today.
- Sentiment Analysis, identify people opinion on a subjective.
- Speech Processing: book a hotel over the phone, TTS (for the blind)
- OCR: both print and handwritten.
- Spelling checkers, grammar checkers, auto-filling, ….. and more

Why NLP is Hard? (1)

- Ambiguity
- Non-standard languages
- Segmentation issues
- Idioms
- Neologisms
- World knowledge
- Tricky entity names
- ……

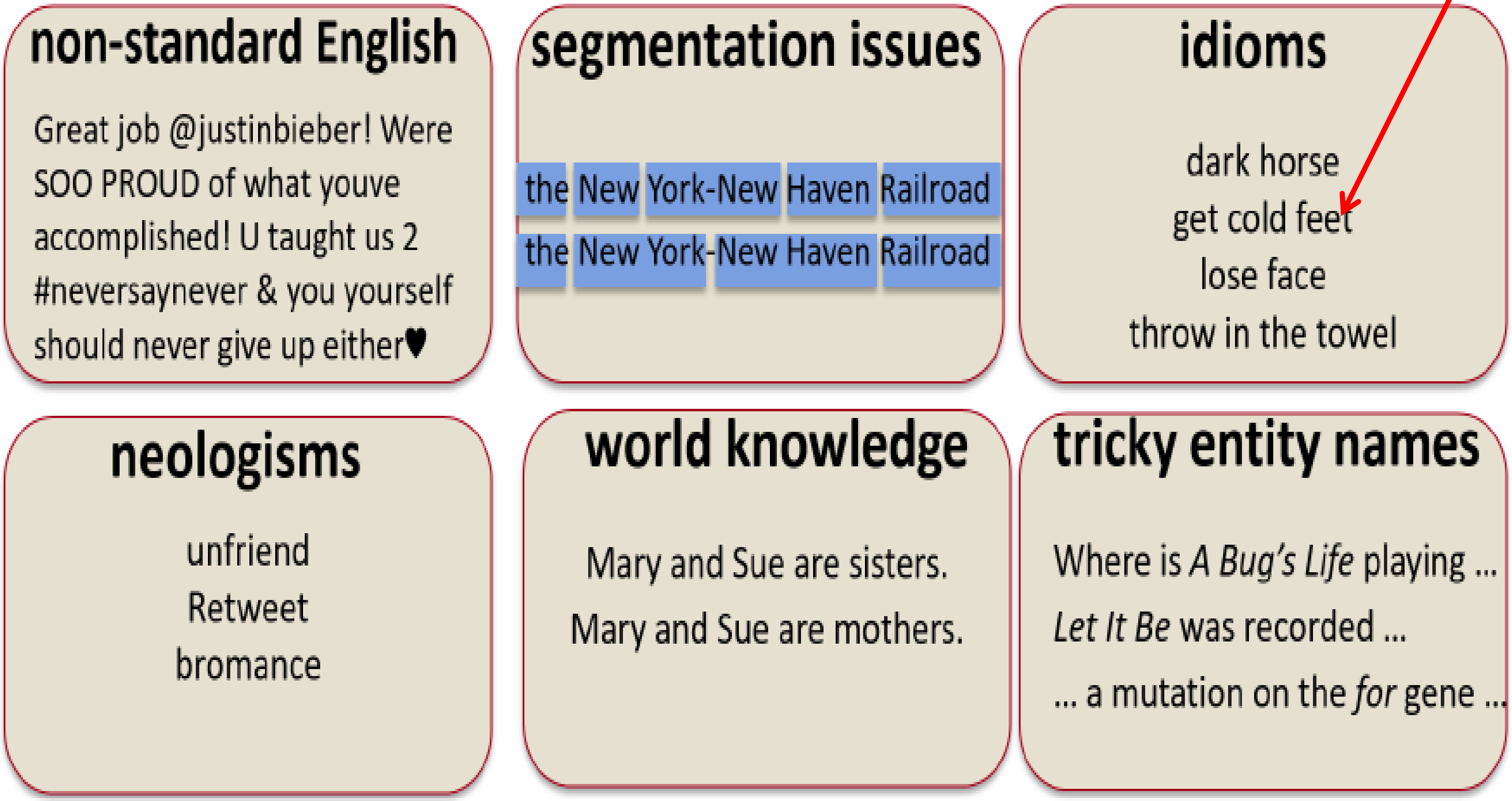
Why NLP is Hard? (2)

- Ambiguity at multiple levels :
  - Word senses: bank (finance or river ?)
  - Part of speech: chair (noun or verb ?)
  - Syntactic structure: I can see a man with a telescope.
  - Multiple: I made her duck.

From Diyi Yang, Georgia Institute of Technology

Why NLP Is Hard? (3)

اشعر بالخوف



non-standard English

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

segmentation issues

the New York-New Haven Railroad

the New York-New Haven Railroad

idioms

dark horse
get cold feet
lose face
throw in the towel

neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

tricky entity names

Where is *A Bug's Life* playing …

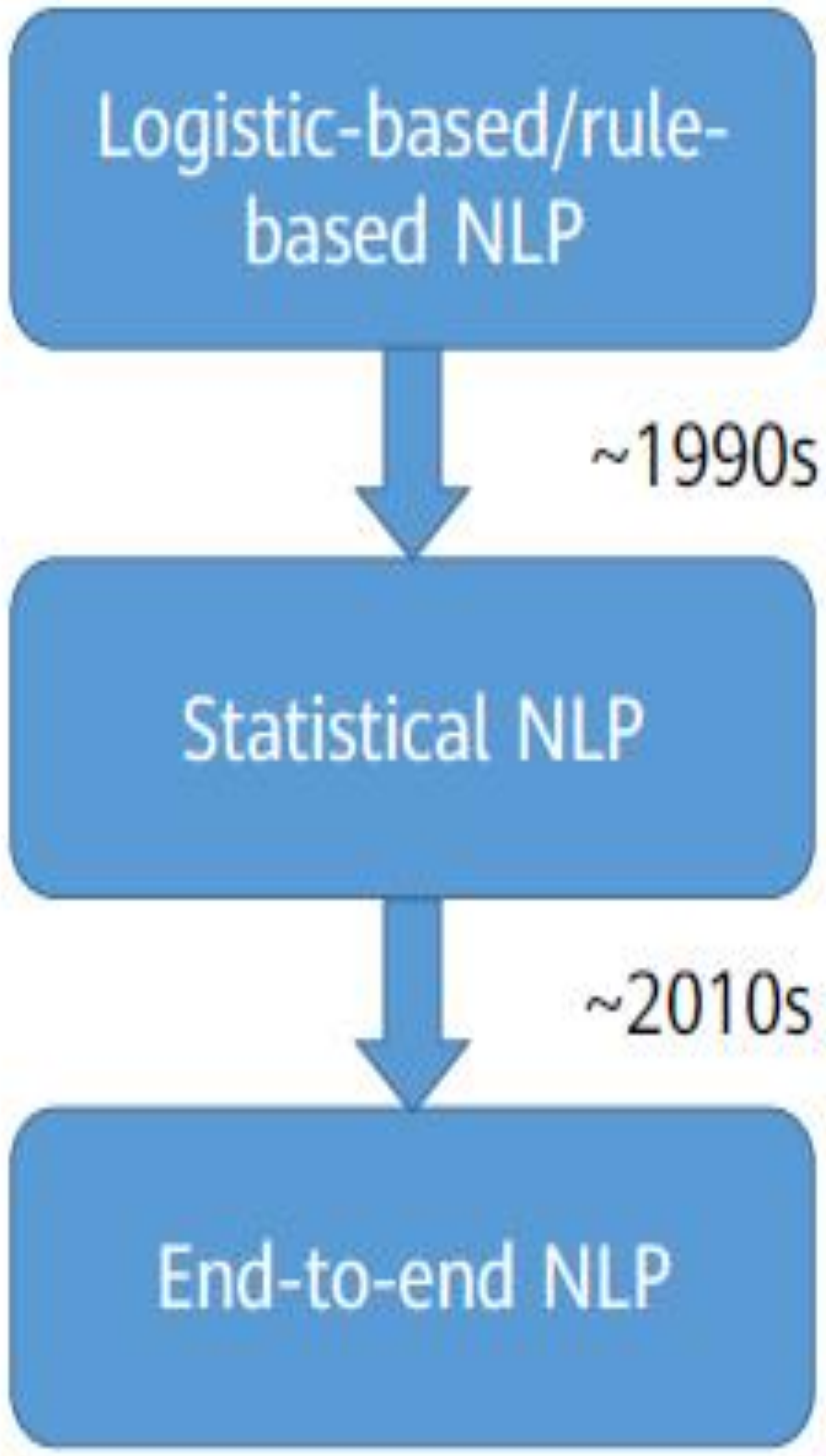
Let It Be was recorded …

… a mutation on the *for* gene …

مصطلحات جديدة

from Dan Jurafsky and Christopher Manning, Stanford University

History of NLP



Logistic-based/rule-based NLP

~1990s

Statistical NLP

~2010s

End-to-end NLP

The traditional NLP pipeline

A (traditional) NLP system may use some or all of the following steps:

Tokenizer/Segmenter

to identify words and sentences

Morphological analyzer/POS-tagger

to identify the part of speech and structure of words

Word sense disambiguation

to identify the meaning of words

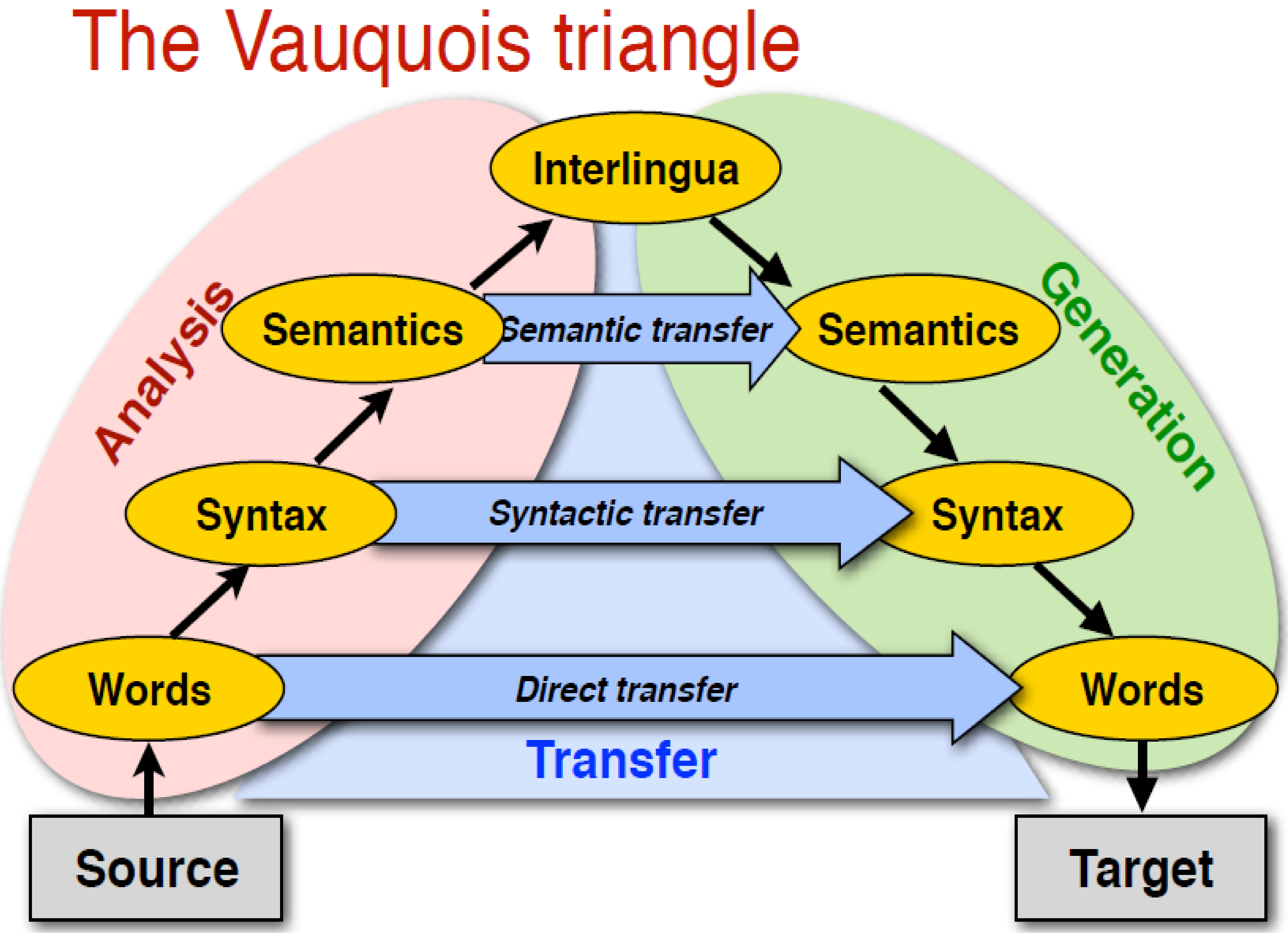
Syntactic/semantic Parser

to obtain the structure and meaning of sentences

Coreference resolution

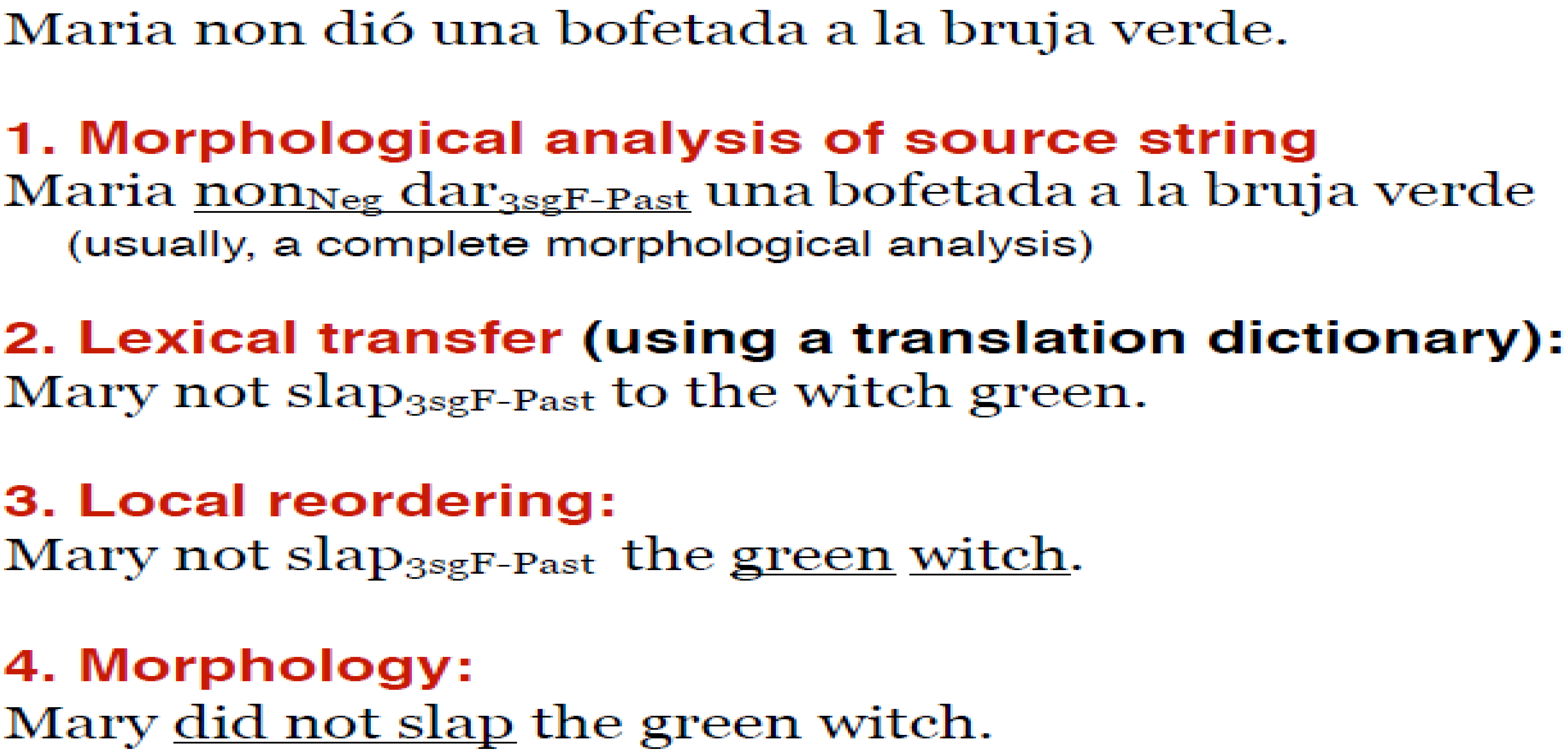
to keep track of the various entities mentioned

Machine Translation



The Direct Translation

- **The Direct Translation approach uses a bilingual dictionary to translate the sentences word by word.**
- **The translation process doesn’t use any intermediate structures, other than some morphological analysis(the task of analyzing the structure and parts of words) or lemmatization(the task of converting a word to its basic dictionary form, for eg, *translating* to *translate*).**



Maria non dió una bofetada a la bruja verde.

1. Morphological analysis of source string

Maria $\underline{\text{non}_{\text{Neg}}\ \text{dar}_{\text{3sgF-Past}}}$ una bofetada a la bruja verde

(usually, a complete morphological analysis)

2. Lexical transfer (using a translation dictionary):

Mary not $\text{slap}_{\text{3sgF-Past}}$ to the witch green.

3. Local reordering:

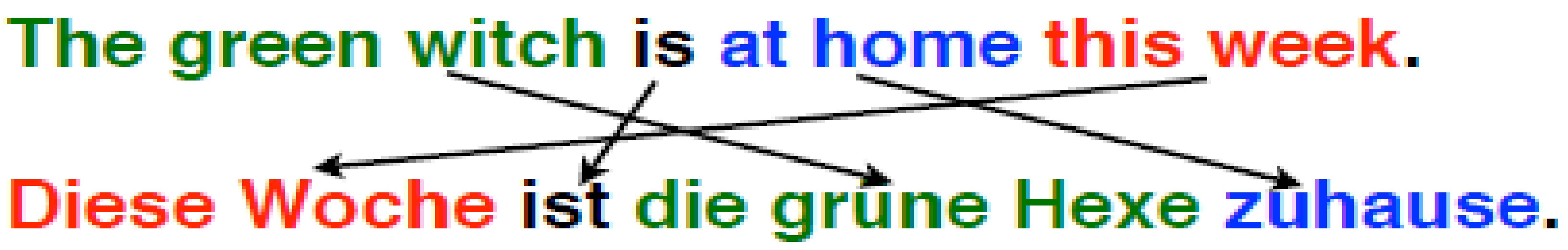
Mary not $\text{slap}_{\text{3sgF-Past}}$ the $\underline{\text{green}}$ $\underline{\text{witch}}$.

4. Morphology:

Mary $\underline{\text{did not slap}}$ the green witch.

Limits of direct translation: Phrasal reordering

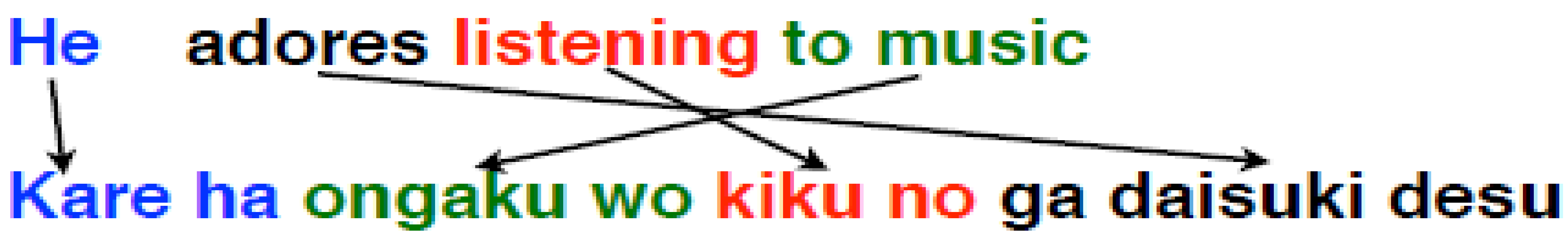
Adverb placement in German:



The green witch is at home this week.

Diese Woche ist die grüne Hexe zuhause.

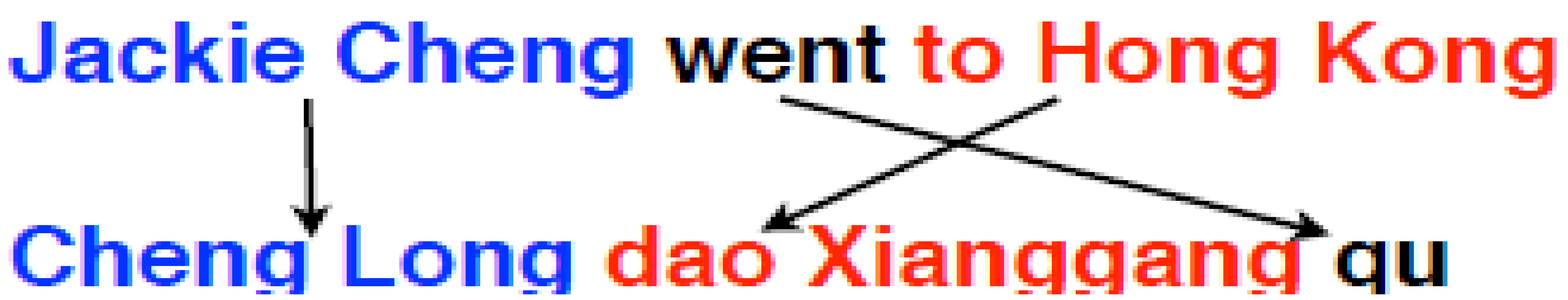
Japanese SOV order:



He adores listening to music

Kare ha ongaku wo kiku no ga daisuki desu

PPs in Chinese:



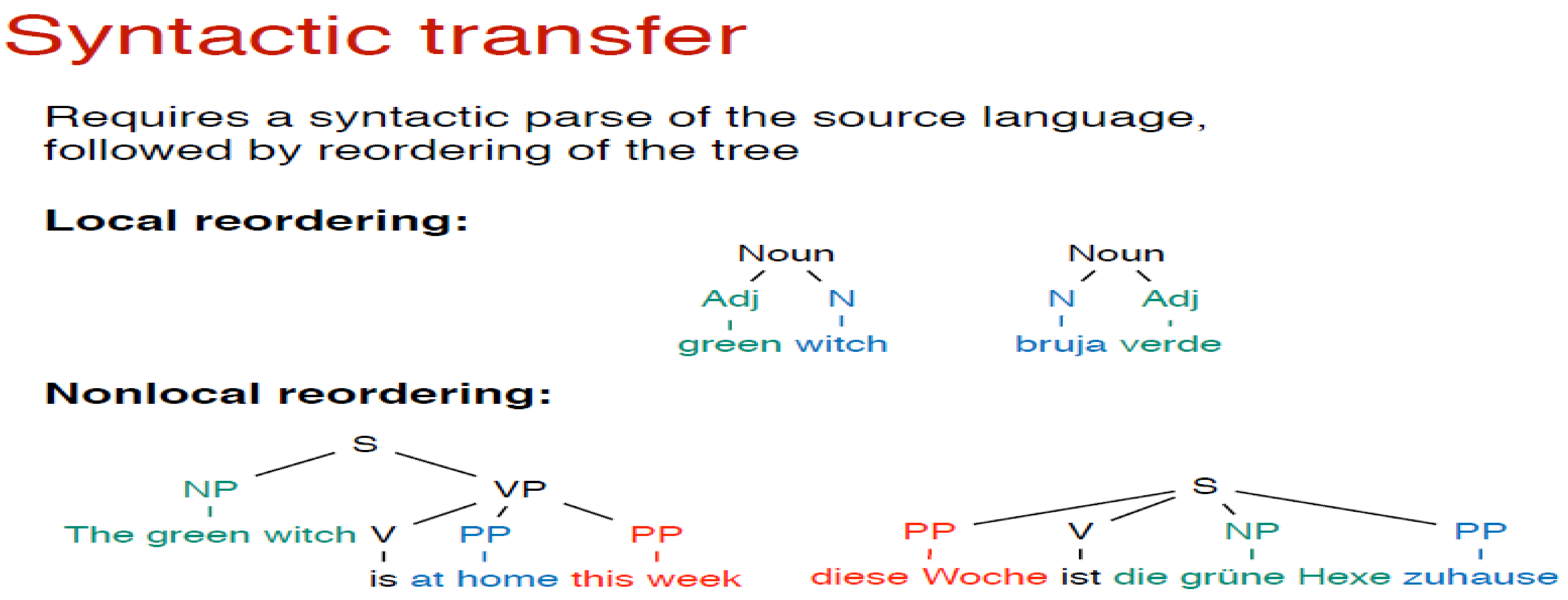
Jackie Cheng went to Hong Kong

Cheng Long dao Xianggang qu

Syntactic Transfer النقل النحوي

- First **parses** حلل the source language sentence to determine its structure.
- **Applies rules to transfer** the resulting structure to the target language parse structure based on the knowledge of differences between the languages.
- **Generates** the target language sentence from this transformed structure.

For similar languages like French and Italian (almost 90% lexically similar), only **syntactic transfer** is needed, where the parsed source structure is transferred to the target parse structure based on syntactic reordering.



Syntactic transfer

Requires a syntactic parse of the source language,
followed by reordering of the tree

Local reordering:

Nonlocal reordering:

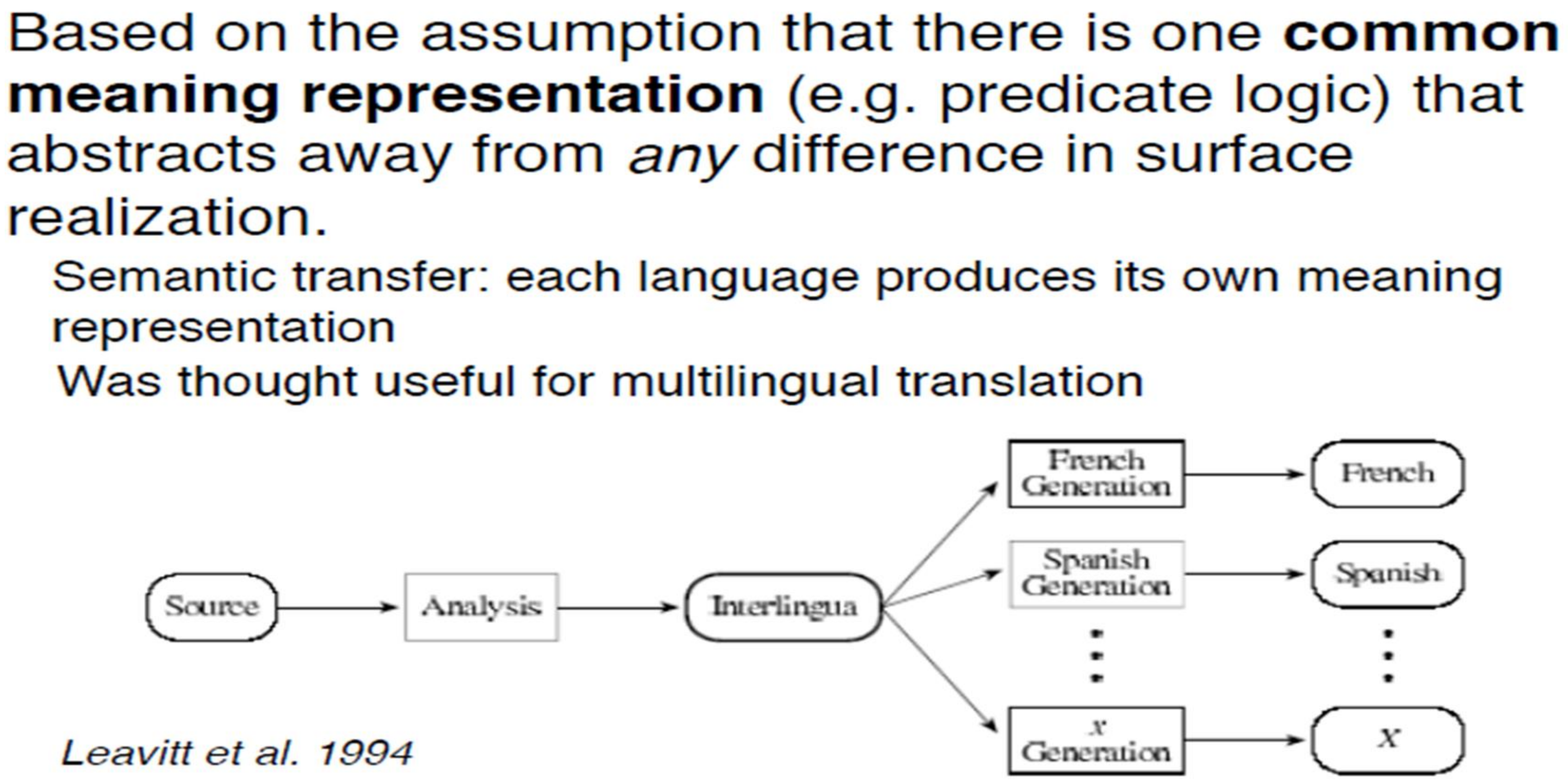
Semantic transfer

For complex translation processes like Korean to English, in addition to syntactic transfer, we also need to do a semantic transfer, where the ** source** structure needs to be transferred according to their meaning and semantic roles in the target sentence

بالإضافة إلى النقل النحوي، نحتاج أيضًا إلى القيام بنقل دلالي، حيث يجب نقل بنية **
المصدر ** وفقًا لمعناها وأدوارها الدلالية في الجملة المستهدفة.

Interlingual approach

The Interlingual approach first analyzes the source language sentence, represents it as an Interlingua, and ** then generates the target language sentence from this interlingua. Interlingua is a language independent representation, that can be based on any representation scheme.**



Based on the assumption that there is one **common meaning representation** (e.g. predicate logic) that abstracts away from *any* difference in surface realization.

Semantic transfer: each language produces its own meaning representation

Was thought useful for multilingual translation

Leavitt et al. 1994

Statistical Machine Translation

We want the best (most likely) [English] translation for the [Chinese] input:

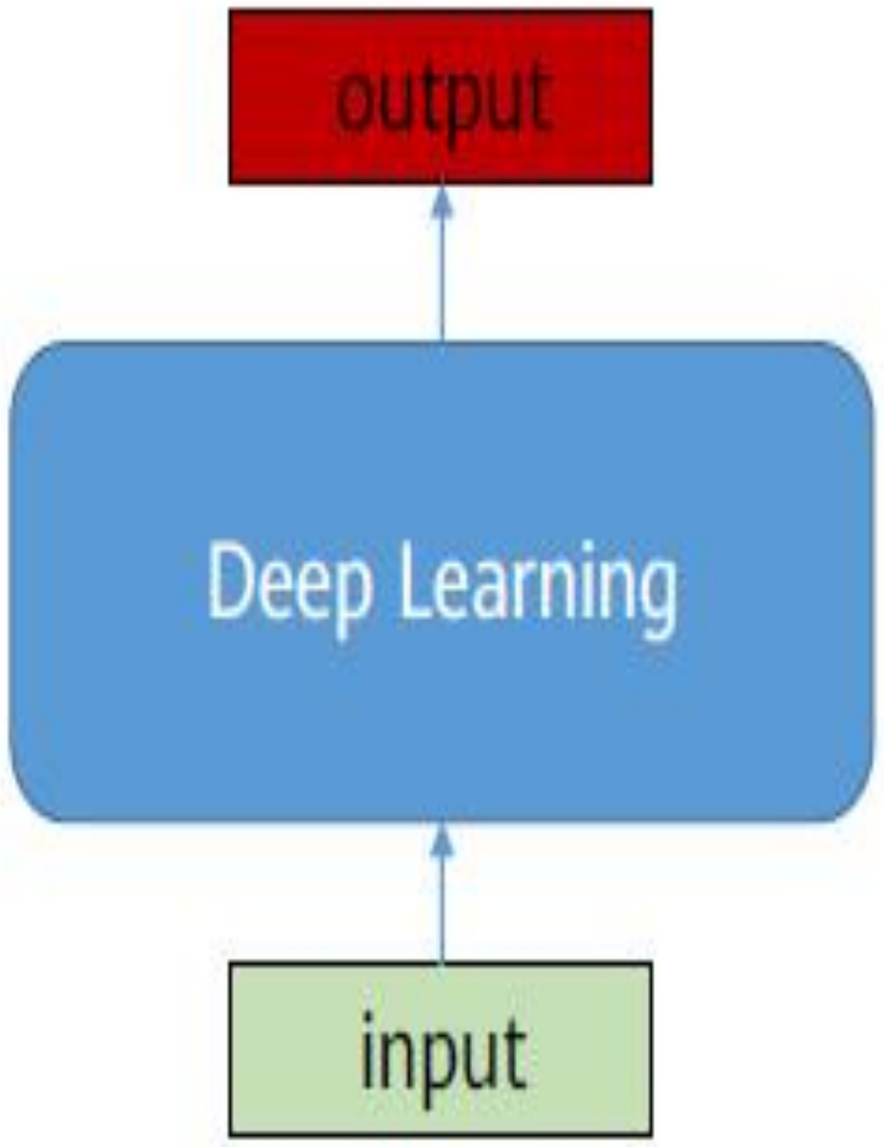
$$\text{argmax}_{\text{English}}\, P(\text{ English} \mid \text{Chinese })$$

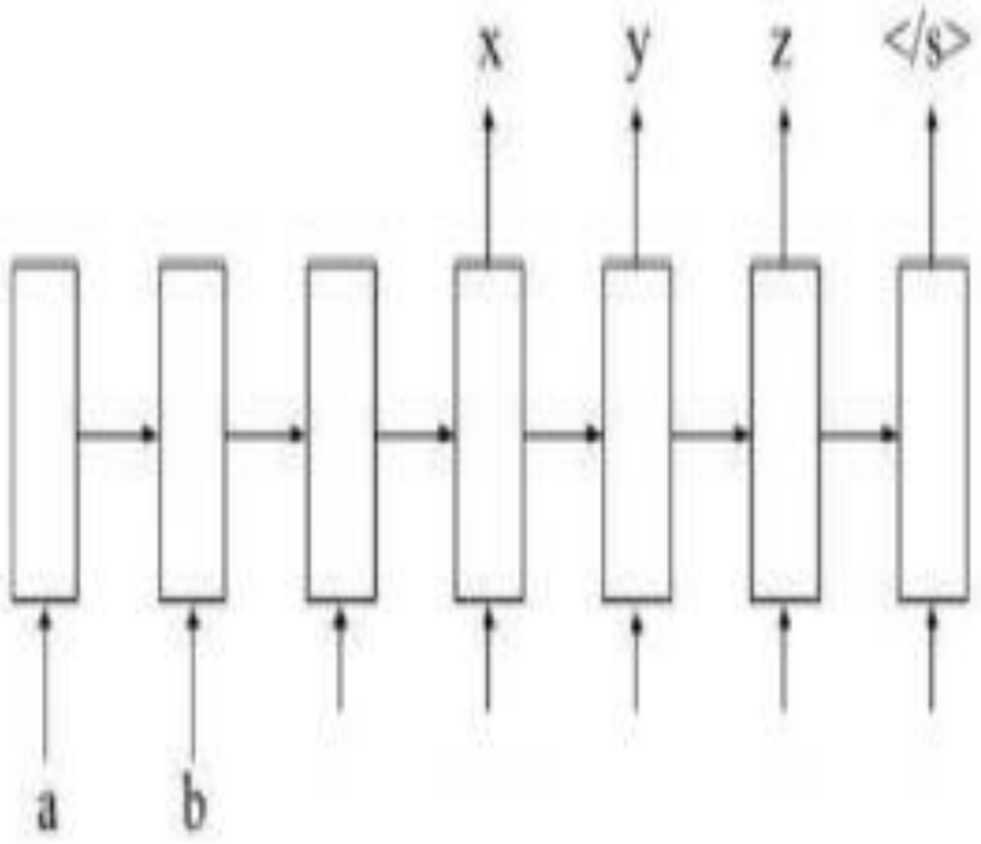
We can either model this probability directly,
or we can apply Bayes Rule.
Using Bayes Rule leads to the "noisy channel" model.

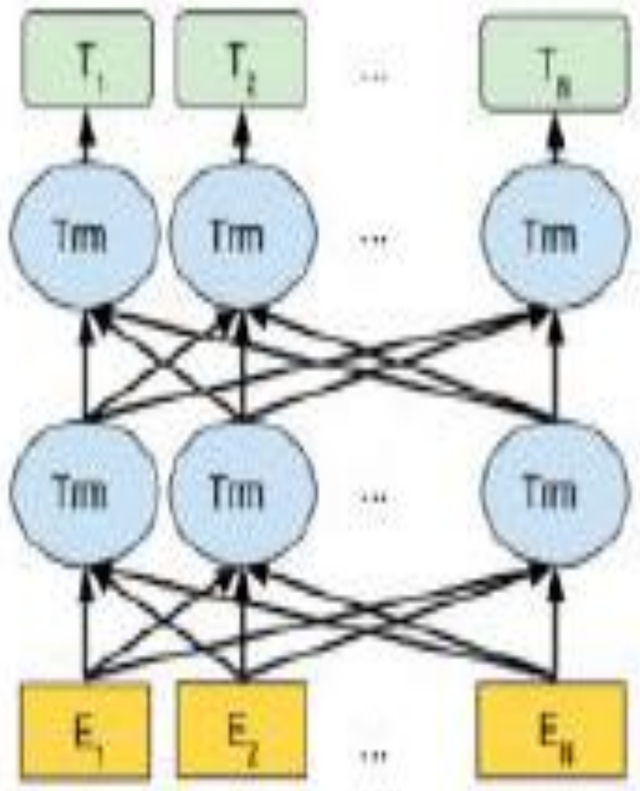
As with sequence labeling, Bayes Rule simplifies the modeling task, so this was the first approach for statistical MT.

Deep Learning and NLP

- Representation Learning (e.g. word embeddings)
- End-to-end Optimization (e.g. NMT)
- Transfer Learning (e.g. BERT)
- Structure Learning (e.g. Transformer)


output
Deep Learning
input




T1
T2
...
TN
Trm
Trm
...
Trm
Trm
Trm
...
Trm
E1
E2
...
EN