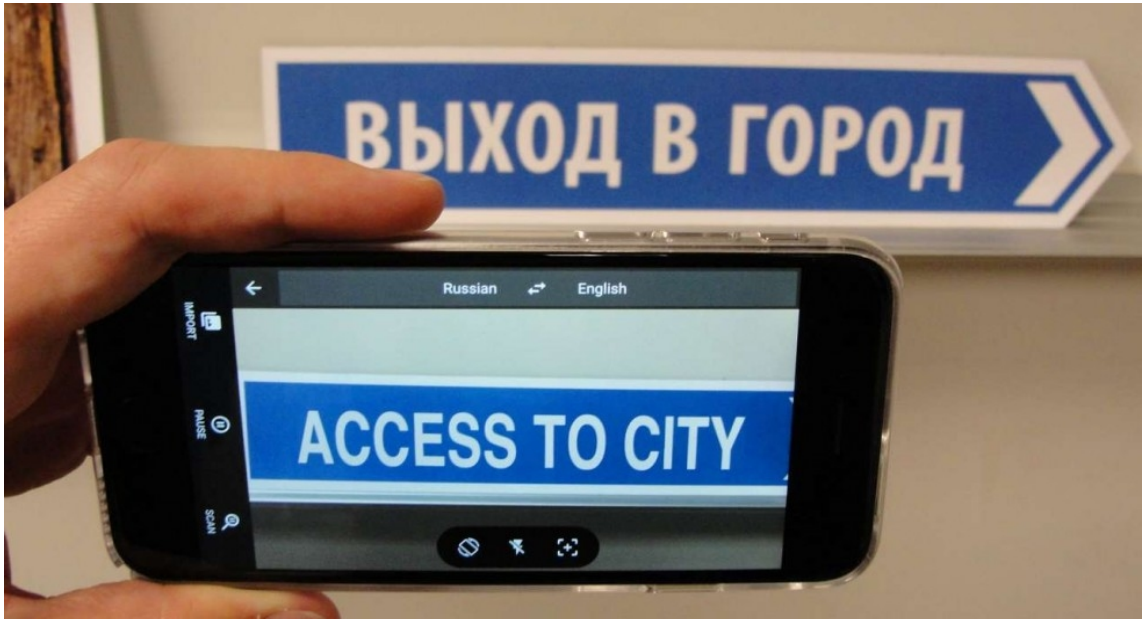
Rule based MT

B. Senthil Kumar
Asst. Professor, CSE
Natural Language Processing

Agenda

- Rule-based MT
- Transfer-based
- Interlingua-based



Rule-based MT

- Parse the source text and produce an intermediate representation – parse tree or abstract representation
- Target text is generated from intermediate representation.
- Contains rules for:
  - Morphology
  - Syntax
  - Lexical selection and transfer
  - Semantic analysis and generation

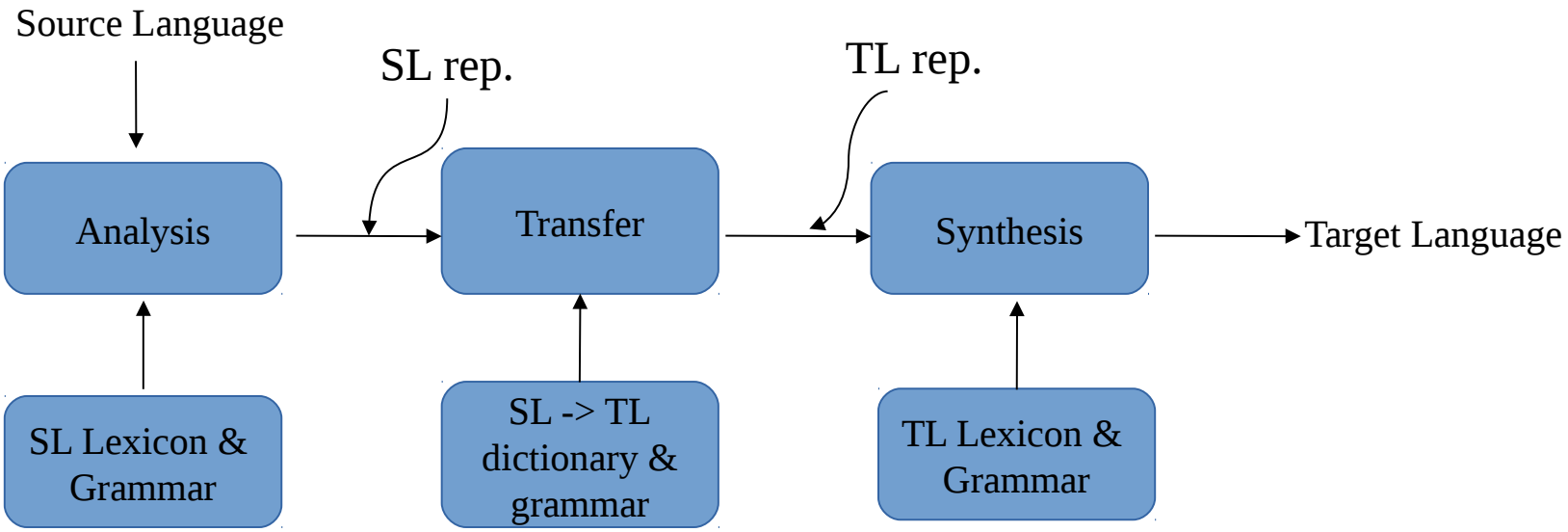
Rule-based categories

- Depending on intermediate representation:
  - Transfer-based MT
  - Interlingua MT

Target Language

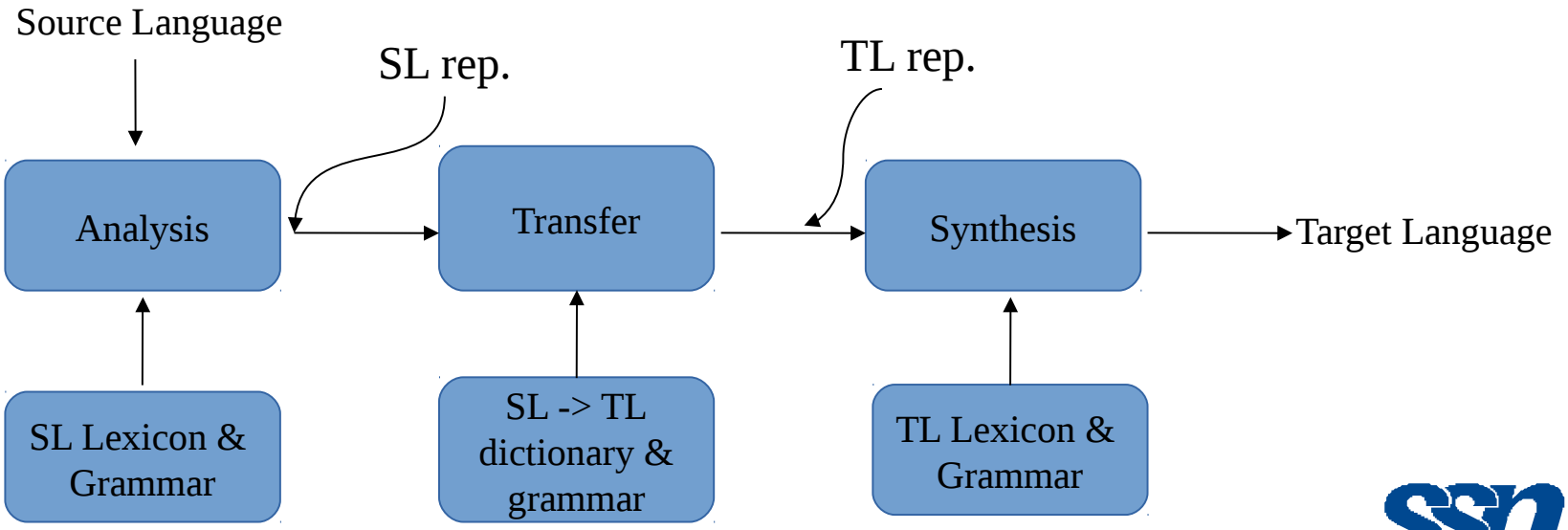
Transfer-based

- Transformation requires understanding of the differences between source and target language
- To get the structure (rep.) -> parsing is required



Transfer-based

- Three components
- Analysis – to produce source language structure.
- Transfer – to transfer the source language representation to a target level representation.
- Generation – to generate target language text using target level structure.



Transfer-based

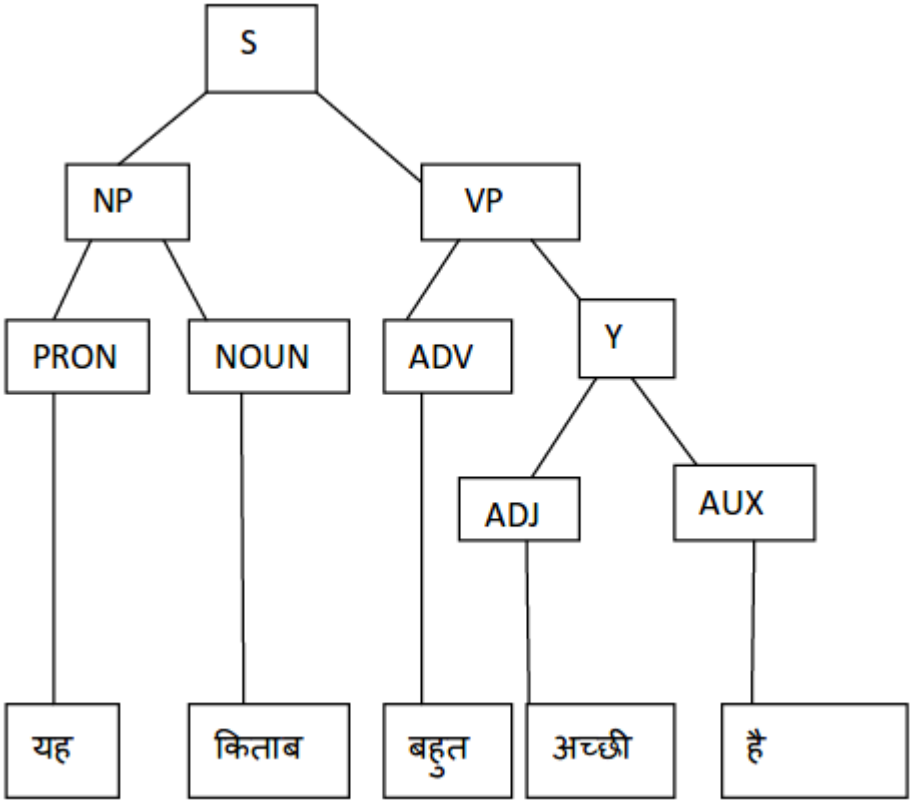
- Source text analysis involves:
  - resolving lexical and syntactic ambiguities.
- Language-pair specific differences are captured in transfer stage.
- Transfer rules are used to transfer from source language representation to target language representation.

Transfer-based

Source: यह किताब बहुत अच्छी है

Target: This book is very good.

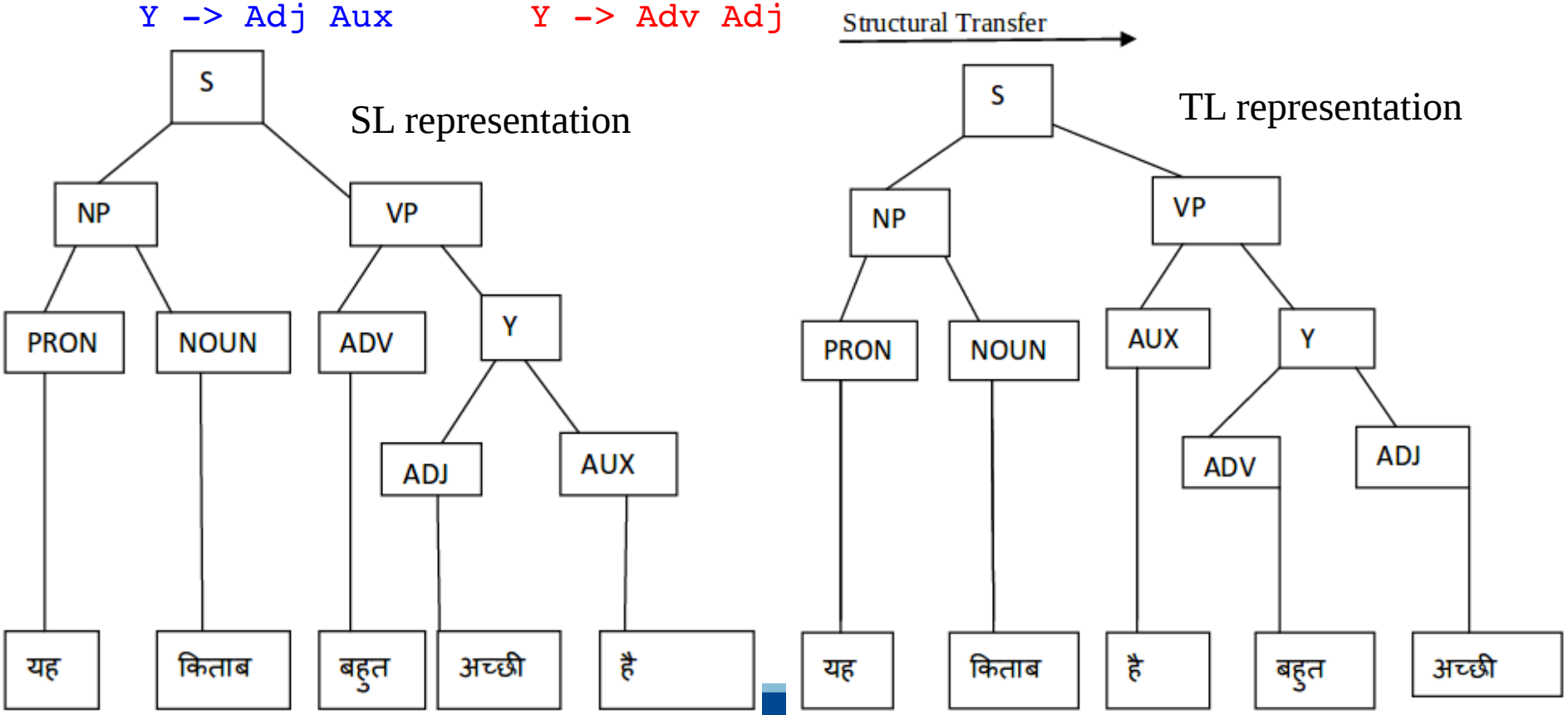
- Consider the source text representation:



Hindi → English

- Transfer rules:

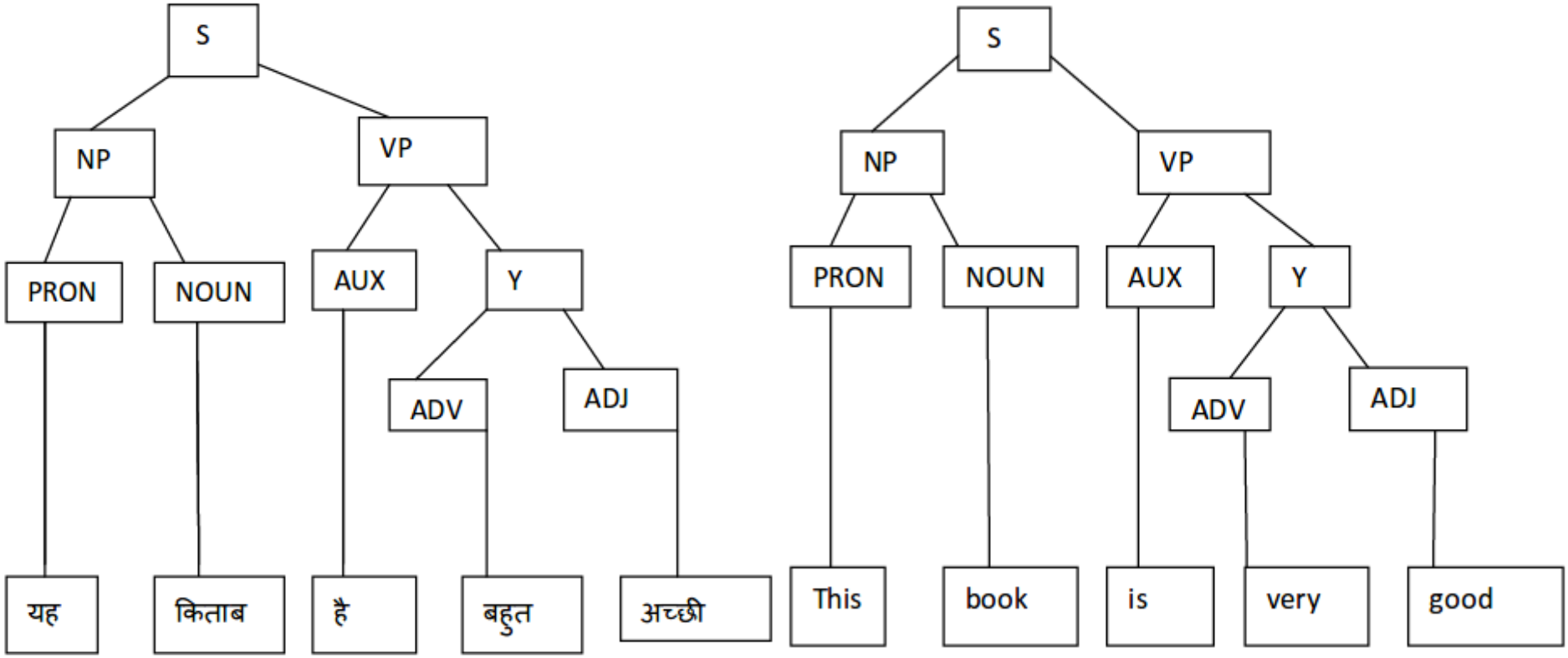
```
S -> NP VP          S -> NP VP
NP -> Pron Noun     N -> Pron Noun
VP -> Adv Y         VP -> Aux Y
Y -> Adj Aux        Y -> Adv Adj
```



Hindi → English

- Generation:

TL representation ⟶ Synthesis

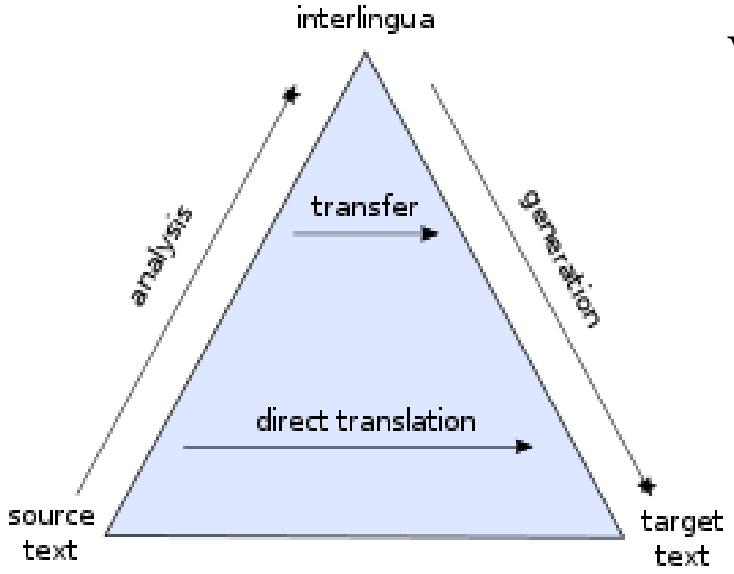


Advantages

- The analysis of source language text (parser) is independent of target language generator.
- Transfer systems easily handle ambiguities that carry over from one language to another.
  - Lexical (POS) and syntactic ambiguities (PP attachment)
- Can be extended to language pairs in multilingual environment.
- For translation among set of languages:
  - An analyzer and generator for each language.
  - A transfer component for a pair of language.

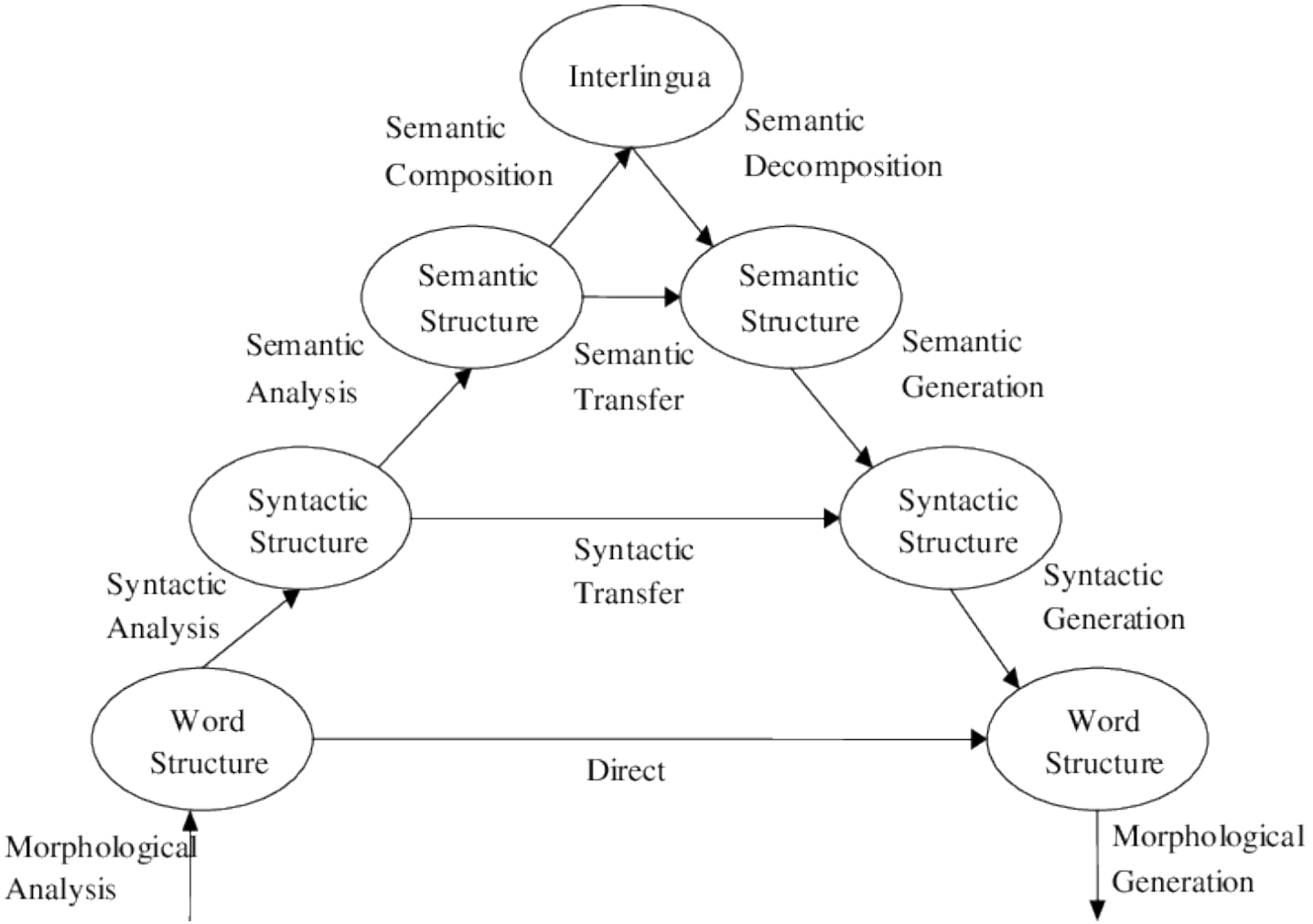
Interlingua-based

- Regardless of varying surface syntactic structure, languages share a common deep structure – Chomsky 1965.
- An interlingua represents all sentences that mean the same thing in the same way regarless of the source language – Jurafsky & Martin 2000.



Vauquois pyramid

Interlingua-based



Interlingua-based

- In the direct approach, words are translated directly without passing through an additional representation.

Source text -> target text

- In the transfer approach the source language is transformed into an abstract, less language-specific representation.
  - Linguistic rules which are specific to the language pair then *transform the source language representation into an abstract target language representation* and from this the target sentence is generated.

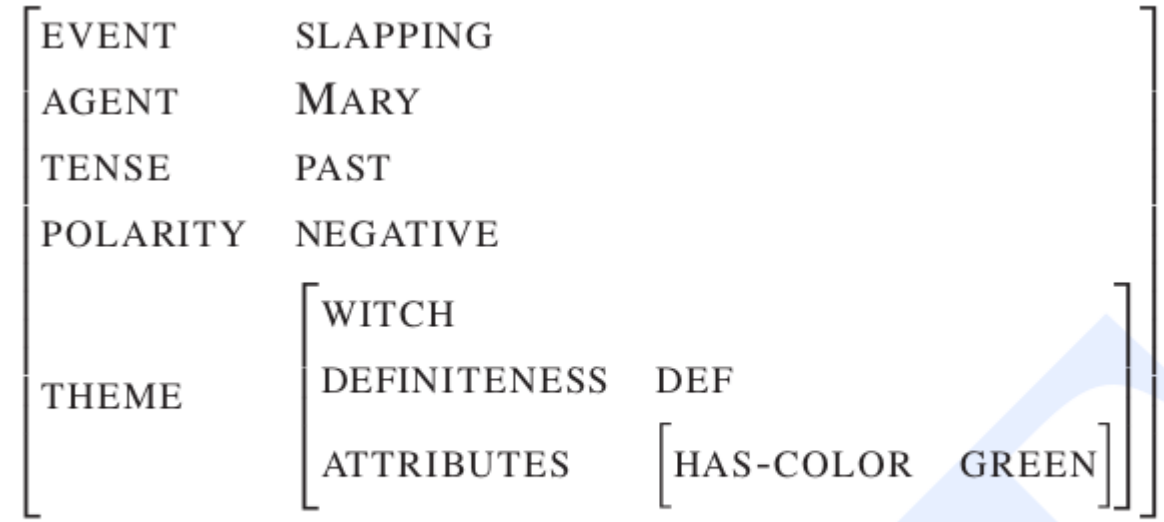
Source text -> source rep -> target rep -> target text

Interlingua ?

- The interlingua can be thought of as a way of describing the *analysis of a text* written in a source language into a target language.
  - morphological, syntactic, semantic (and even pragmatic) characteristics, that is "meaning"
- Single underlying representation for both SL and TL.
- Abstracts away from language-specific characteristics.
- Creates a “language-neutral” representation.
- Can be used as a “pivot” representation in the translation.

Interlingual representation

- Interlingual representation as unificationstyle feature structure.



Interlingual representation of *Mary did not slap the green witch.*

Interlingua-based

- The source language is transformed into an interlingua, i.e., an abstract *language-independent* representation.
- The target language is then generated from the interlingua.
- KANT – CMU, 1989:
- The source text is processed using the source language grammar and lexicon to produce a Source F-Structure.
- In the Interpretation stage, mapping rules map lexical items onto semantic concepts, and syntactic arguments onto semantic roles, forming the intermediate representation.

Interlingua-based

- The interlingua representation comprises information from all necessary levels of linguistic analysis; *lexical, syntactic, semantic and pragmatic.*
- In the generation stage, target mapping rules indicate how the interlingua representation maps onto the appropriate Target F-Structure.

Interlingua-based

- Interlingual representation

"The primary power supply component will supply the necessary 240 Volts DC to the input lead." ⟹

```
(*a-supply
   (tense future)
   (mood declarative)
   (punctuation period)
   (source (*o-power-supply-component
              (reference definite)
              (number singular)
              (attribute (*p-primary))))
   (theme (*u-volt-dc
             (reference definite)
             (number plural)
             (attribute (*p-necessary))
             (quantity
                (*c-decimal-number
                (integer "240")
                (number-type cardinal)
                (number-form numeric)))))
   (goal_to (*o-input-lead
                (reference definite)
                (number singular))))
```

Interlingua-based

- Interlingual representation

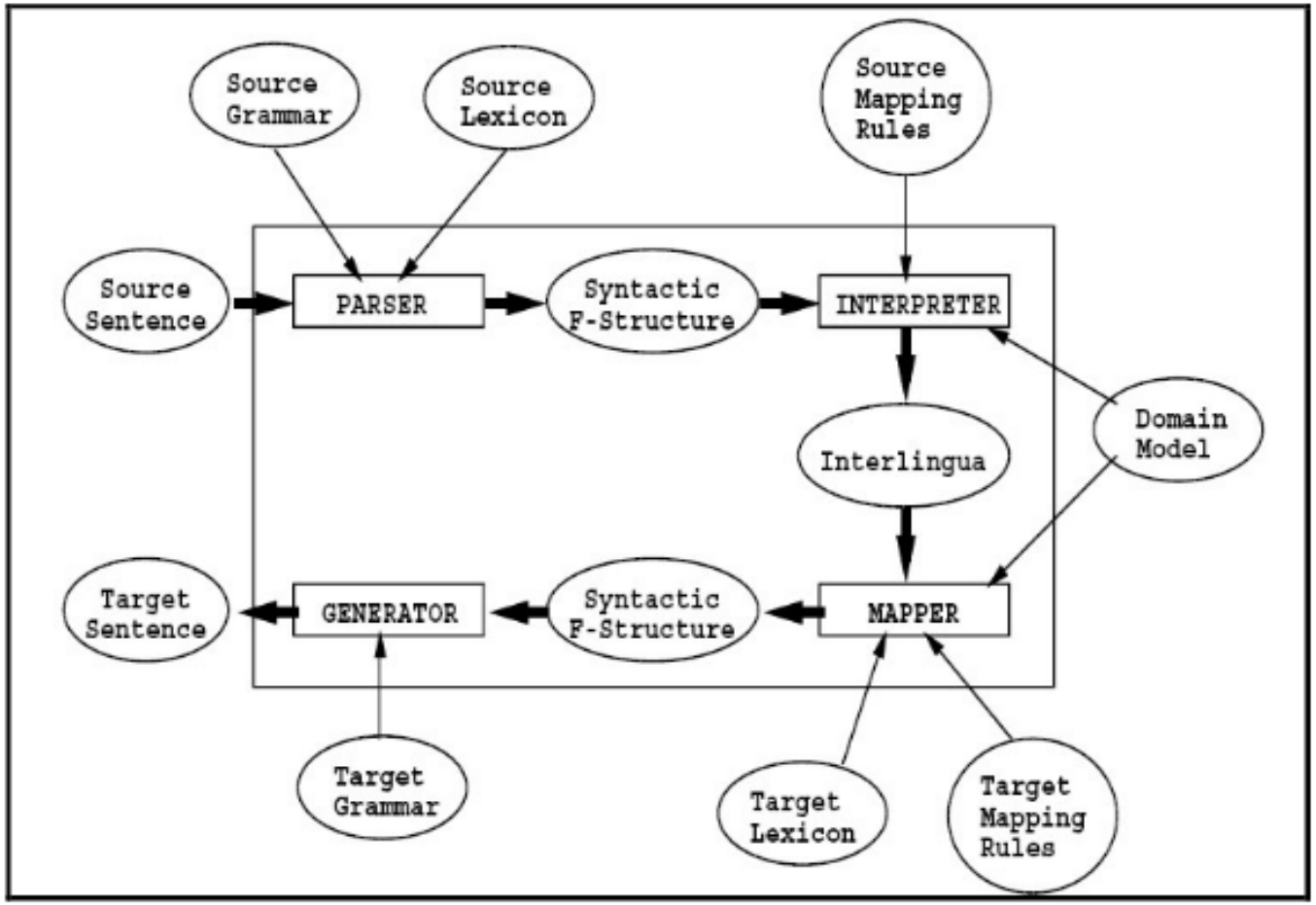


Figure 9: The run-time architecture of KANT

References

- *Natural Language Processing and Information Retrieval*, Tanveer Siddiqui, Tiwari, Oxford
- *Speech and Language Processing*, Daniel Jurafsky, Martin, Pearson, 2006.
- *Interlingua-based Machine Translation Systems: UNL versus Other Interlinguas*, Sameh Alansary