

NLP/MT
Principles

EBMT Principles
and Solution

EBMT & Rule-based
MT

EBMT & Knowledge-
based MT

EBMT & Stat.;
Evaluation

The logo of the University of Hamburg, featuring the letters 'U' and 'H' in a stylized, white, sans-serif font on a light red background.

Rule based Machine Translation

Cristina Vertan

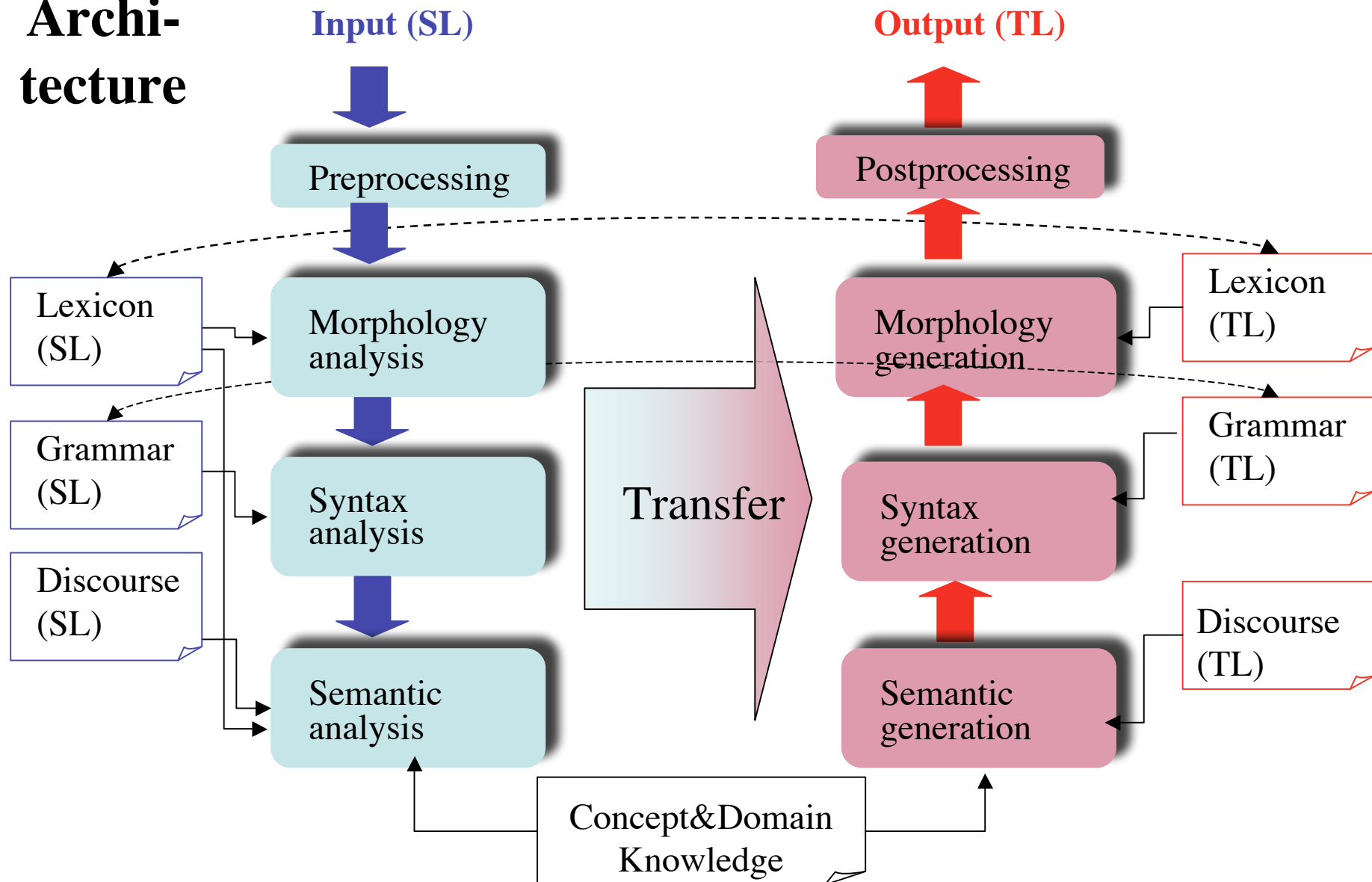
University of Hamburg • Informatics Department

Natural Language Systems Group

WWW: <http://nats-www.informatik.uni-hamburg.de/~cri/>

E-Mail: vertan@informatik.uni-hamburg.de

Architecture



Knowledge Sources

- Bilingual (Multilingual) Lexicon
- Thesauri
- Grammars for SL and TL
- World /Domain Knowledge base
- Discourse memory

Bilingual (Multilingual) Lexicon Example

French Lexicon

```
<entry id=„123“>  
  <word> pomme </word>  
  <PoS> Noun </PoS>  
  <genus> F </genus>  
  <number> sg. </number>  
  <case> N,AG,D</case>  
  <transl.> ref. 576 E</transl>  
</entry>
```

English Lexicon

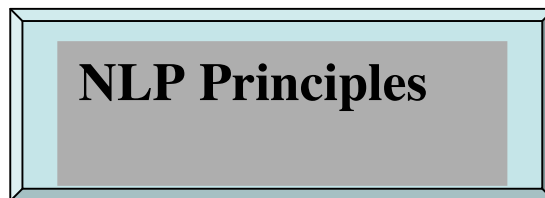
```
<entry id=„576“>  
  <word> apple </word>  
  <PoS> Noun </PoS>  
  <genus> </genus>  
  <number> sg. </number>  
  <case> N,A,G,D</case>  
  <transl.> ref. 123 S </transl>  
</entry>
```

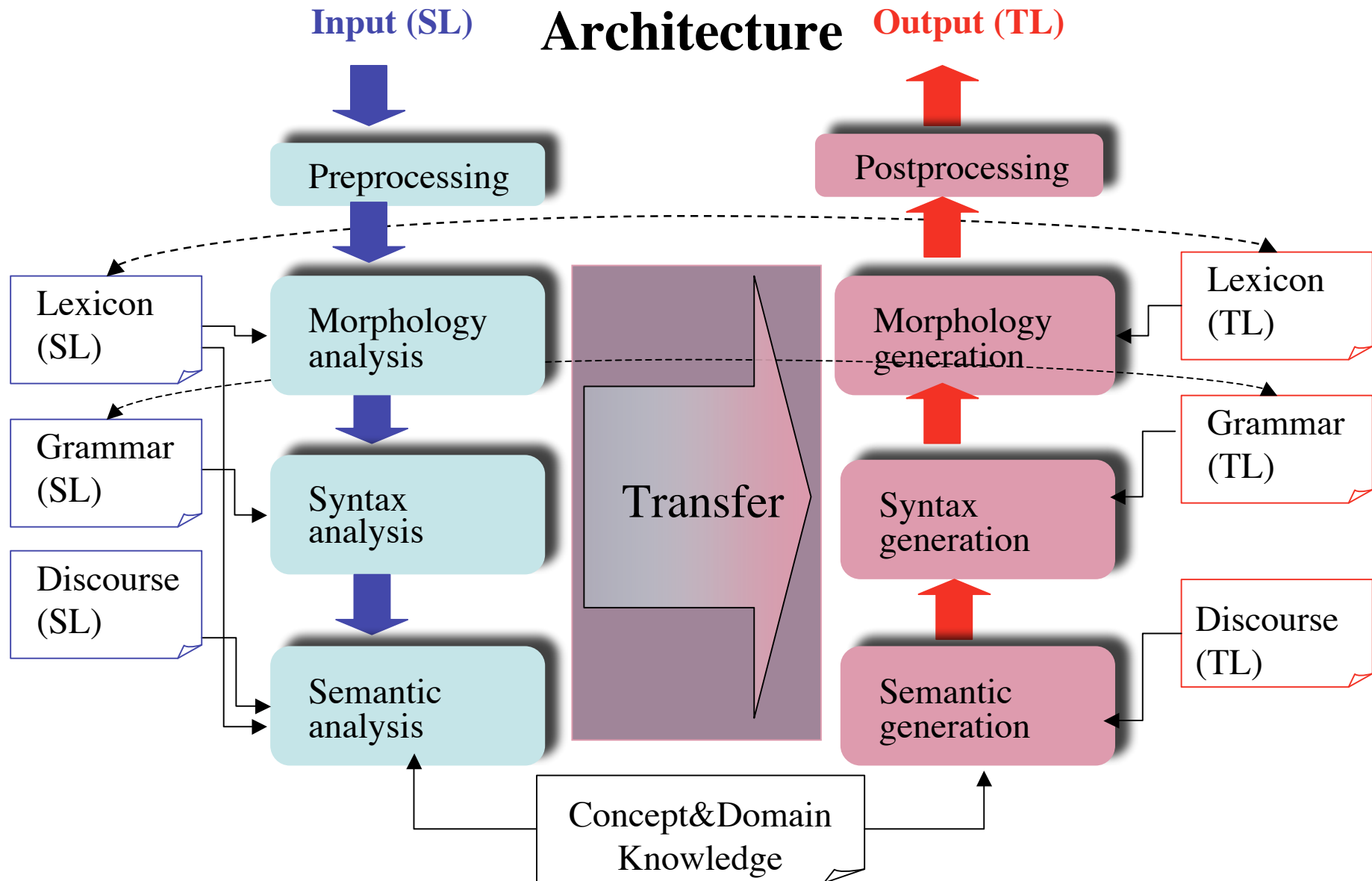
Thesauri

- Are a particular form of lexicons and contain fixed expressions and their translations
- Expressions contained in such thesauri are replaced from the very beginning by their translations, and are no longer object of syntactic or semantic interpretations
- E.g.:
 - *United States = Statele Unite*
 - *Civil law = Cod Civil*
- Sometimes abbreviations are also part of thesauri:
- E.g.:
 - *Dvs. = Dumneavoastră = You (politeness)*
- Thesauri are domain specific

Morphological, Syntactic and Semantic Analysis

Follow general principles in NLP





Transfer levels

- Lexical level - corresponds to direct (word to word) systems
- Structural transfer :
 - Syntactic level - classical transfer systems
 - Semantic level - interlingua systems.

Morphological analysis -1-

Germ: Ich wollte die Äpfel gestern kaufen

Fr: Je ? la ? hier acheter

les

Informations about:

Inflection: Apfel (sg. masc) inflection class N23

wollen (present)

Declination: Äpfel (acusative, pl.)

Conjugation: Ich wollte, etc.

Stem lexicon

Langenscheidts
Universal-
Wörterbuch

French

Ich = Je

die = la, les

gestern = hier

kaufen = acheter

Morphology-representation

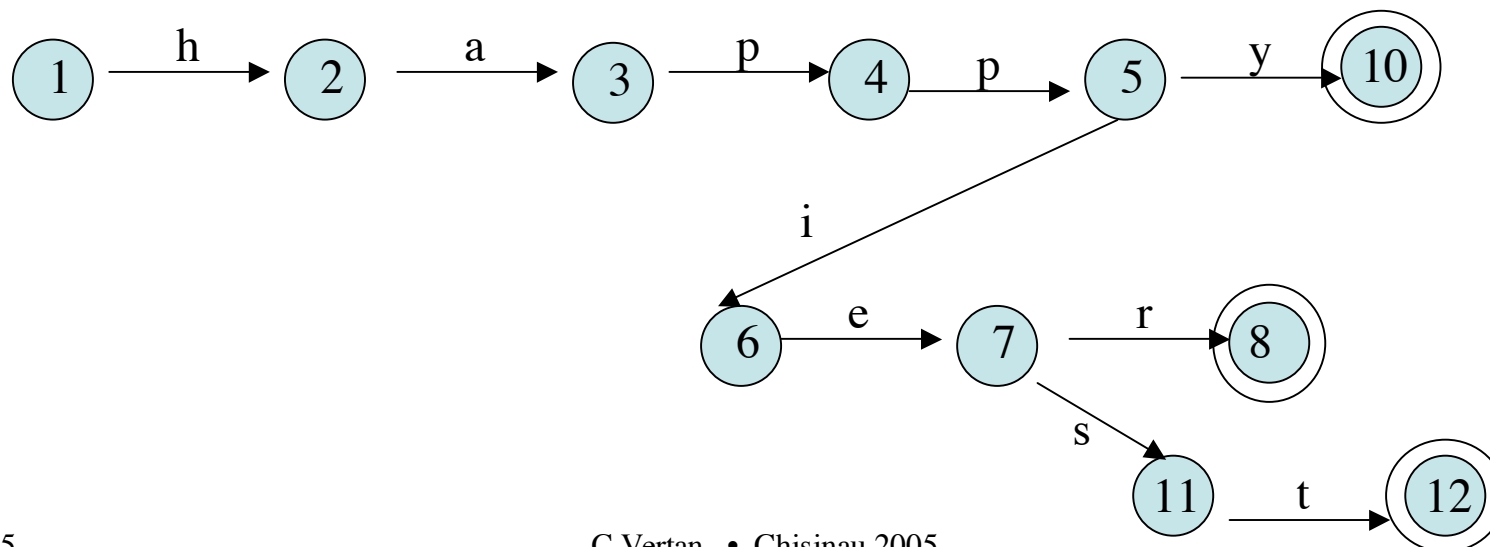
- Rules:

(lex=V, cat=v,+finite, person=3rd, number=sing, tense=pres) \leftrightarrow V+s

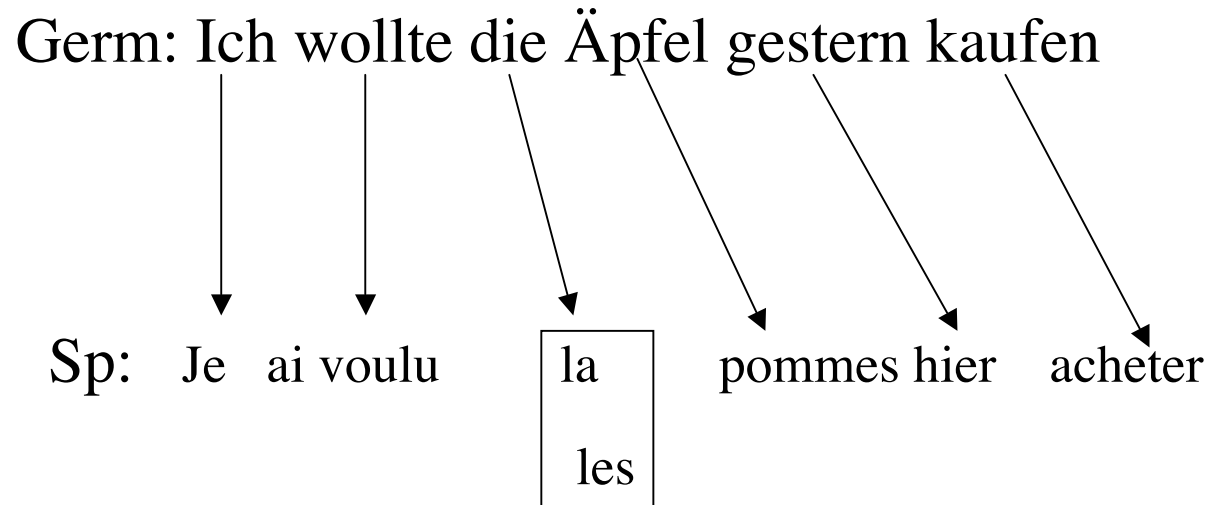
Exception

(lex=be, cat=v,+finite, person=3rd, number=sing, tense=pres) \leftrightarrow is

- Finite State Transducers (FST)



Morphological analysis -2-



Full-form lexicon

Langenscheidts
Universal-
Wörterbuch

French

Ich = Je

wollte = voulu

die = la, les

Äpfel = pommes

gestern = hier

kaufen = acheter

Limitations of morphological analysis -1-

- From the previous example: after the morphological analysis the translation would be:
 - *J 'ai voulu la/les pommes hier acheter*
- 2 Problems:
 - no correct word-order
 - Ambiguity when translating „die“
- The word-order can be solved by introducing transfer rules: e.g. the verb has to be moved from the last position (according to the German order) near the auxiliary (according to the French order). But not all such changes can be defined by rules.

Limitations of morphological analysis -2-

- Lexical Ambiguity:
 - Categorial ambiguity: the same word can belong to more than one PoS E.g. *last* (engl.):
 - *Verb*: The show *lasts* 2 hours
 - *Adjective*: *last* time
 - *Adverb*: He is the *last*
 - Homography and Polysemy (the same word has more meanings) e.g. Bank (engl.) capital (sp.)
 - Translation ambiguity: e.g. the English *leg* can be translated in Spanish with *pierna* (*human*), *pata* (*animal, table*), *pie* (*chair*), *etapa* (*of a journey*)
- Structural ambiguity : *la pommes* or *les pommes*, or complicated syntactical problems

Lexical transfer

- Consists usually of:
 - Replacement of lexical elements in SL by their correspondents in TL
 - If necessary correction of word-order
e.g. Adj Noun (engl) → Noun Adj (sp.)
 - Without problems when:
 - There is a translation equivalent in TL
 - Many to one translation i.e. more lexical items in SL are translated by the same item in TL
- E.g. *pegar, chocar, acertar* and *golpear* (sp.) can be all translated with *hit* (engl.), i.e the system doesn't have to analyse the semantic context of the verb.

Lexical transfer - Problems -1-

- One-to many translation (one word in SL has different translations according to the context)
 - E.g. *Wall* (engl.) will be translated by *muro* (sp.) *Mauer* (germ.) if it is outside and *pared* (sp.), *Wand*(germ.) inside. In this case the semantic features of the lexical entries, have to be compared with the input sentence. E.g. for the translation of “know” the grammatical context must be known :
 - I know him* (engl.) → *Ich kenne ihn* (germ.)
 - I know a solution* (engl.) → *Ich weiß eine Lösung* (germ.)

Lexical transfer - Problems -2-

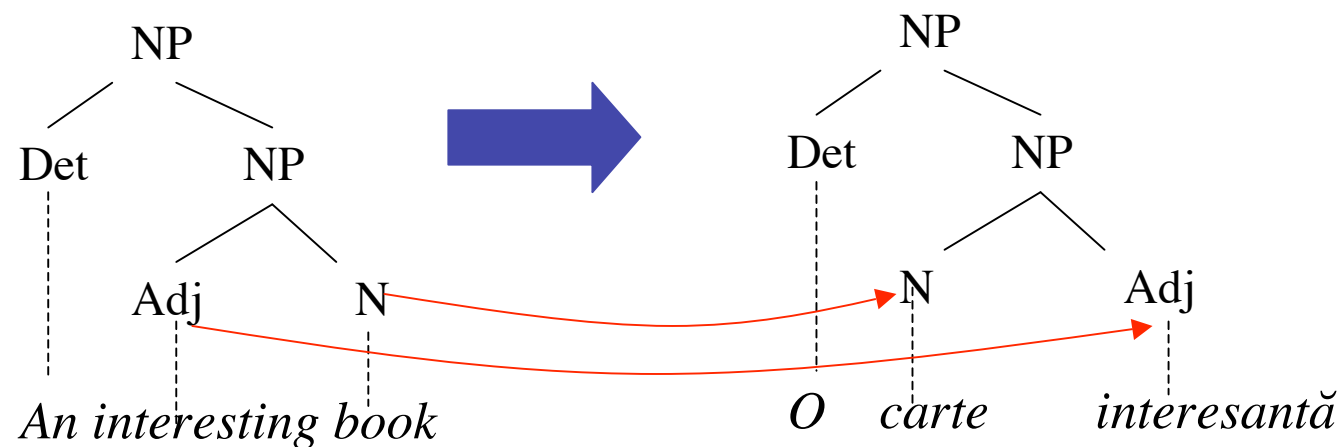
- Different translations depending on domain-specific/world knowledge
 - E.g. *bibliotcă* (rom.) is translated into German by:
 - *Bibliothek* if it belongs to an academic institution or is private
 - *Bücherei* if it is a public library.
- Lexical gaps - single-word concepts in one language which can only be rendered by two or more words in the other
 - E.g. *madrugó* (sp.) = *got up early* (engl.)
 - This cannot be solved only by lexical transfer because in the English lexicon there is no entry “*get up early*”
 - Lexical gaps must not be confused with specific cultural words (e.g. *tapas*, *alcázar* (sp.), *quiche* (fr.), *sarmale*(rom)) which are usually not replaced in the translation process

Structural transfer

- is always necessary if the structure in SL can not be transferred to the TL, or it does not fit exactly due to semantic or stylistic rules
- The deeper the analysis the more differences between languages disappear from the representation
- Solution : transfer rules

Syntactic transfer -1-

- Mapping between the surface structure of sentences: a collection of tree-to-tree transformations is applied recursively to the tree of the SL sentence in order to construct a TL tree
- The simplest form correspond to word-order rearrangements in lexical transfer:

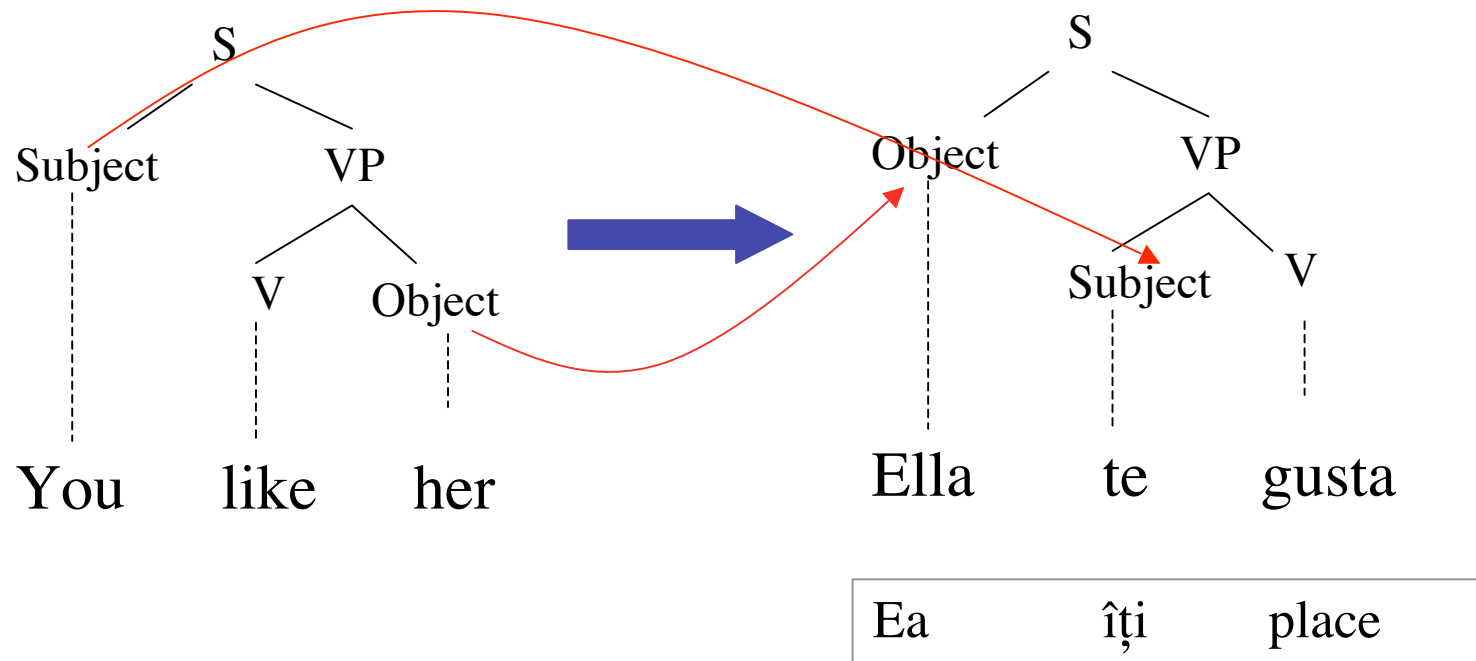


Syntactic transfer -2-

- Tree-to-tree transformations:
 - Recursive
 - Top down process
 - One side of the tree-to-tree transfer rule is matched against the input structure, resulting in the structure on the right-hand side
- Rules have to cover not only such simple cases but also:
 - Thematic divergences
 - Head switching
 - Structural differences
 - Lexical gaps
 - Lexicalization
 - Categorical divergences
 - Collocational divergences

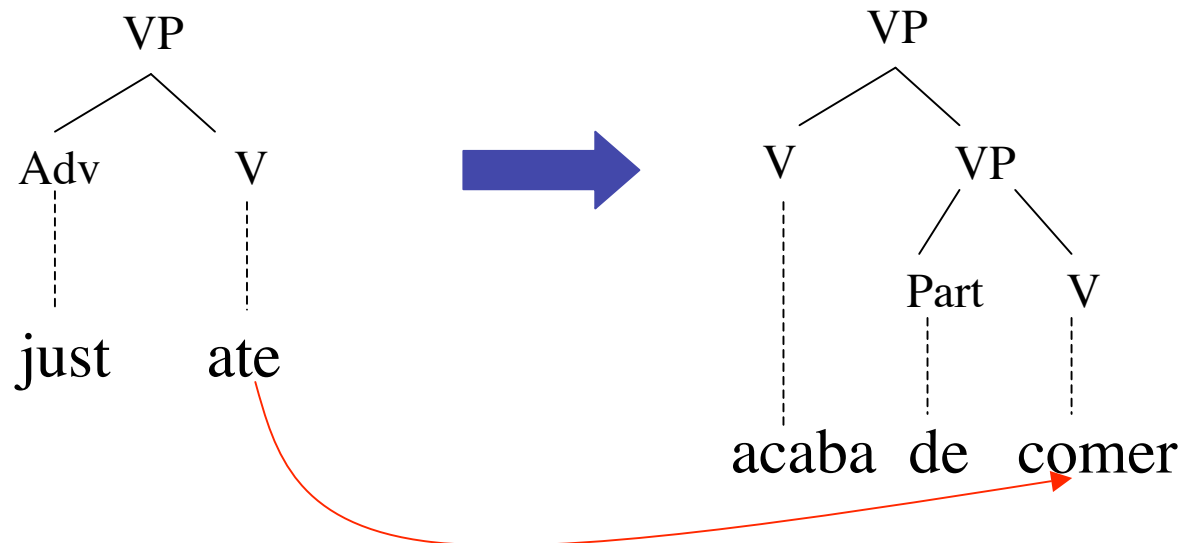
Syntactic Transfer - Thematic divergences-

- Thematic divergences refer to changes in the grammatical role played by arguments of a predicate



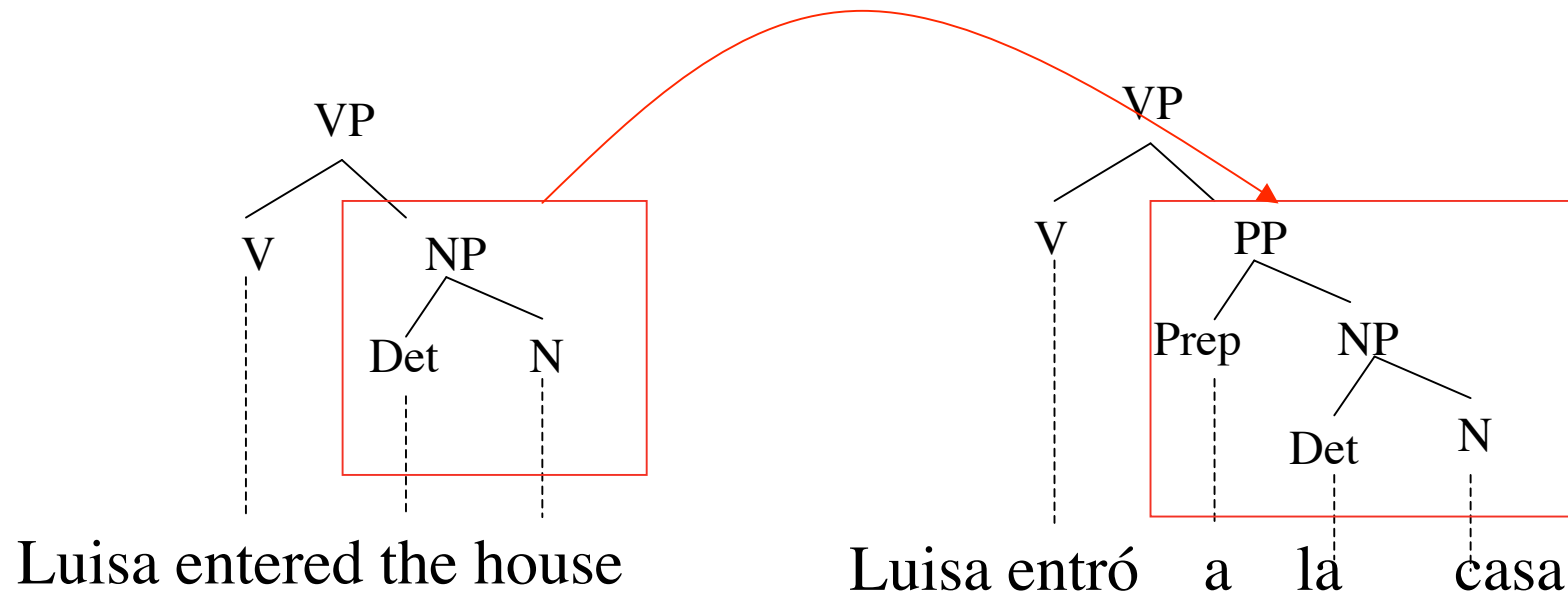
Syntactic Transfer - Head Switching-

- The syntactic head of an expression in one language is translated as modifier, a complement or some other constituent in an other language



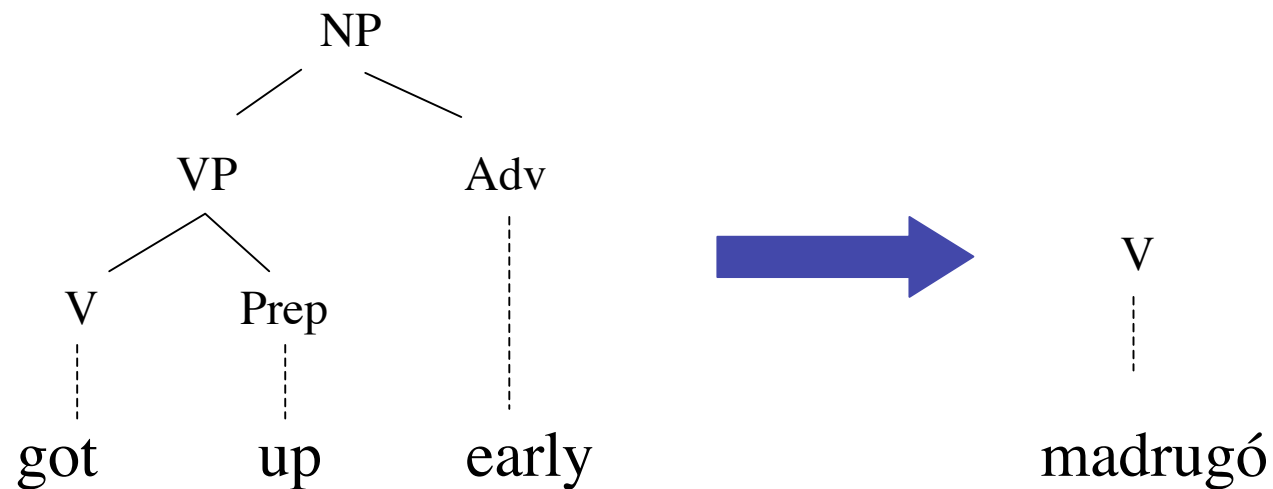
Syntactic Transfer - Structural divergence

- Different sub-constituents for the same constituent.



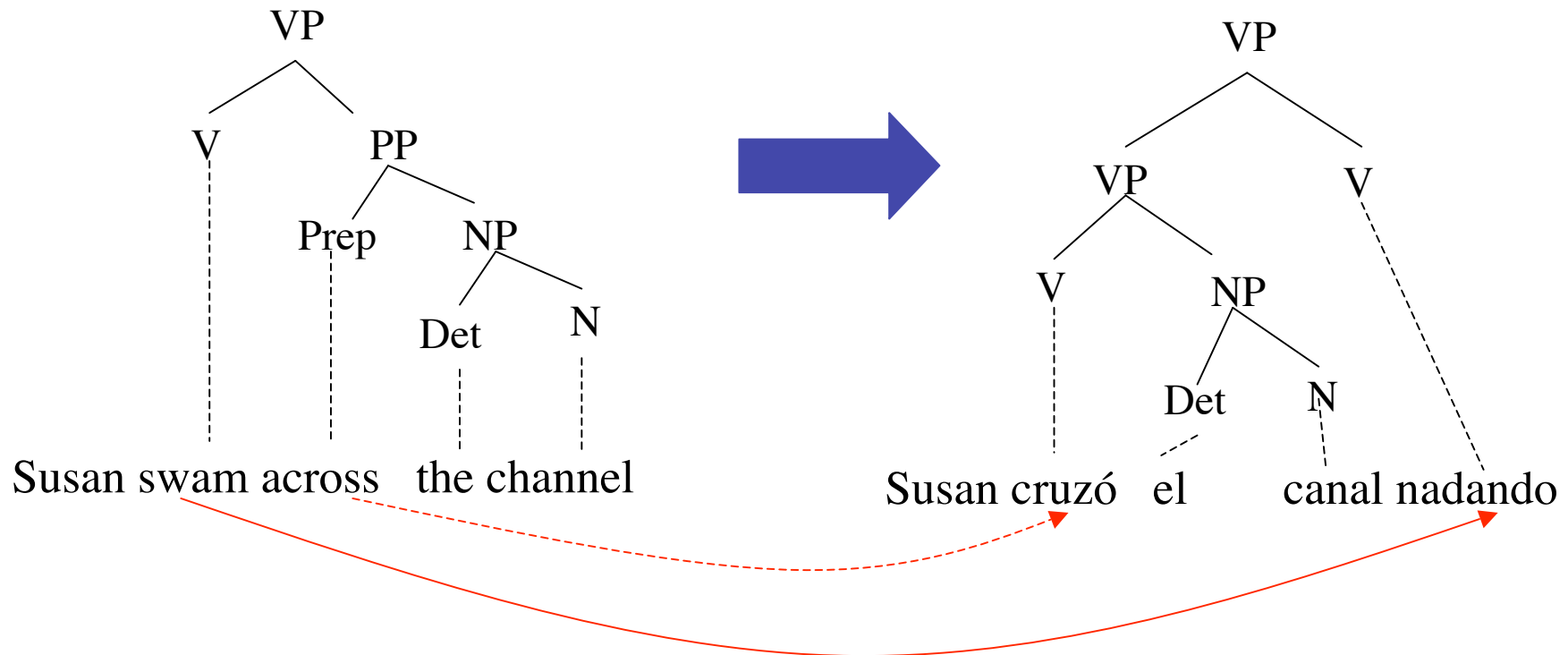
Syntactic Transfer - Lexical gaps-

- For such cases special rules must be provided



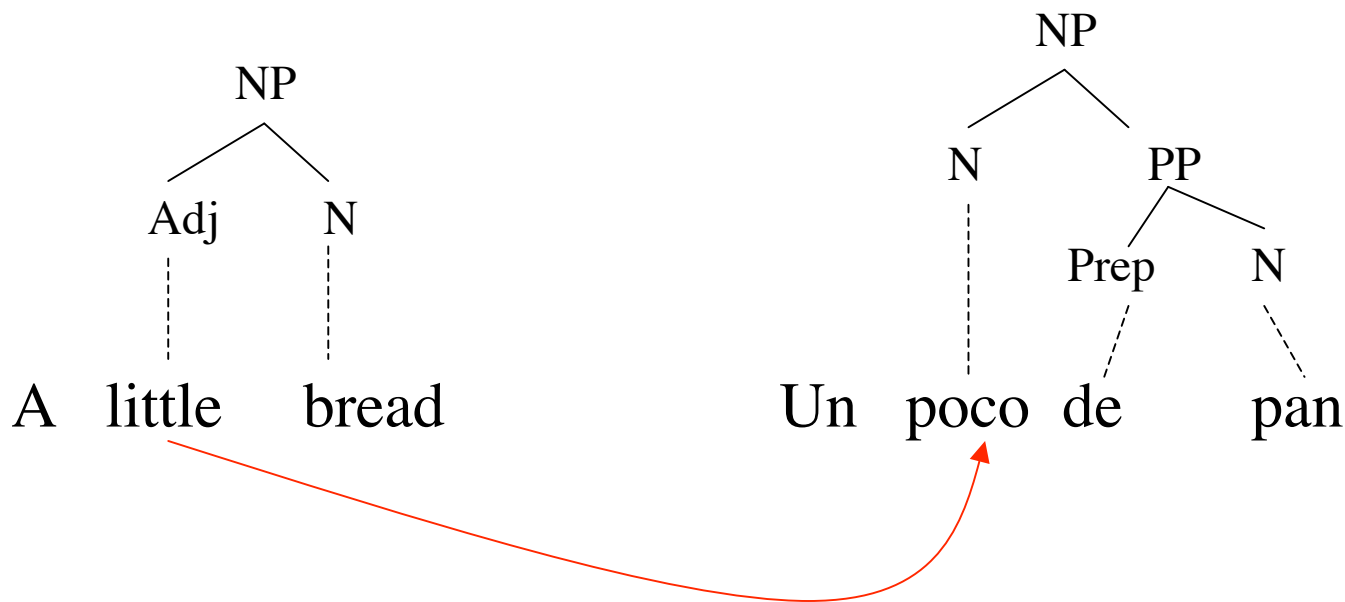
Syntactic Transfer - Lexicalization-

- Languages distribute semantic content differently within a sentence



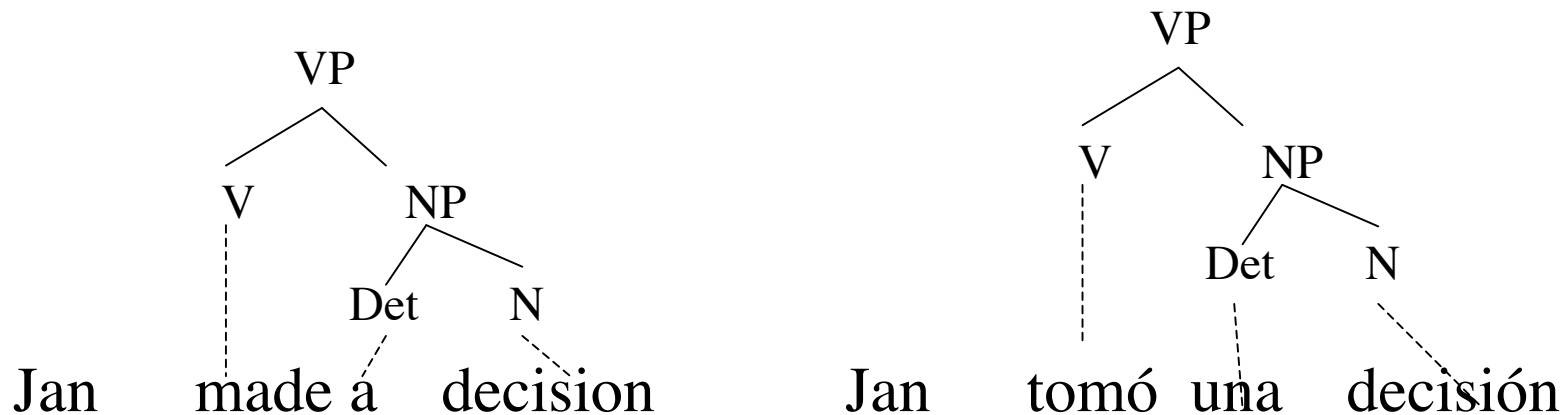
Syntactic transfer - Categorical divergence-

- Although a one-to-one-translation exists, some words must be rendered via different syntactic categories. Sometimes this involves also head switching (*I am hungry* (engl.) → *Tengo hambre* (sp.))



Syntactic Transfer - Collocational divergence-

- arise when the modifier, complement or head of a word is different from its default translation.



Solution: list all combinations of relevant collocations and insert specific rules for each (*take a walk = dar una caminata, be thirsty = tener sed* etc.)

Semantic Transfer

- In syntactic transfer many transformations are only variations on each other
- Long dependencies are difficult to handle
- Semantic transfer interprets translation as a relation between language-dependent representations.
- The transformations are also recursive but they apply on the semantic representation
- There are many semantic representation formalisms
 - E.g. QLF (Quasi-Logical form) based on predicate logic

Semantic Transfer - Example

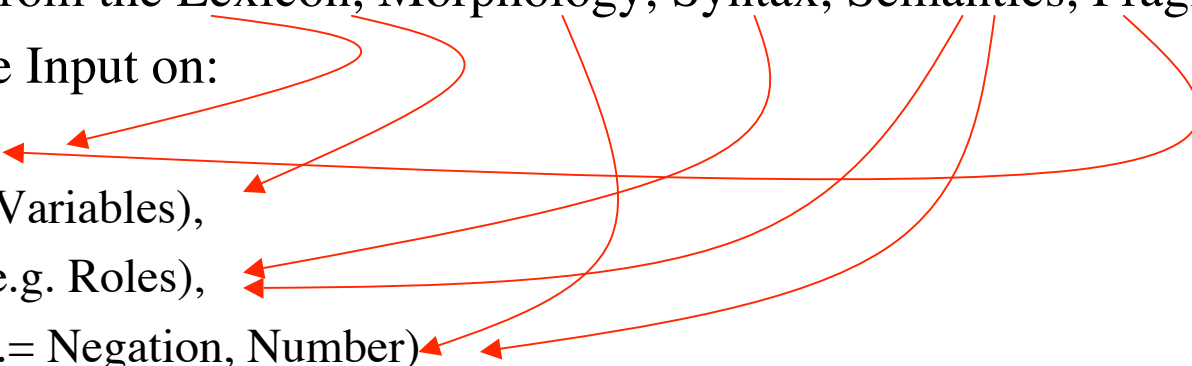
- Engl.: the girl slept \rightarrow (Sp.) la chica durmió
[past, FormE] \Leftrightarrow [past, FormS] :- FormE \Leftrightarrow FormS.
[sleep, S, ArgE] \Leftrightarrow [dormir,S,argS]:- ArgE \Leftrightarrow ArgS.
qterm(the,X,RestE) \Leftrightarrow qterm(def,X, RestS):-RestE \Leftrightarrow RestS.
[girl,Y] \Leftrightarrow [chica,Y]

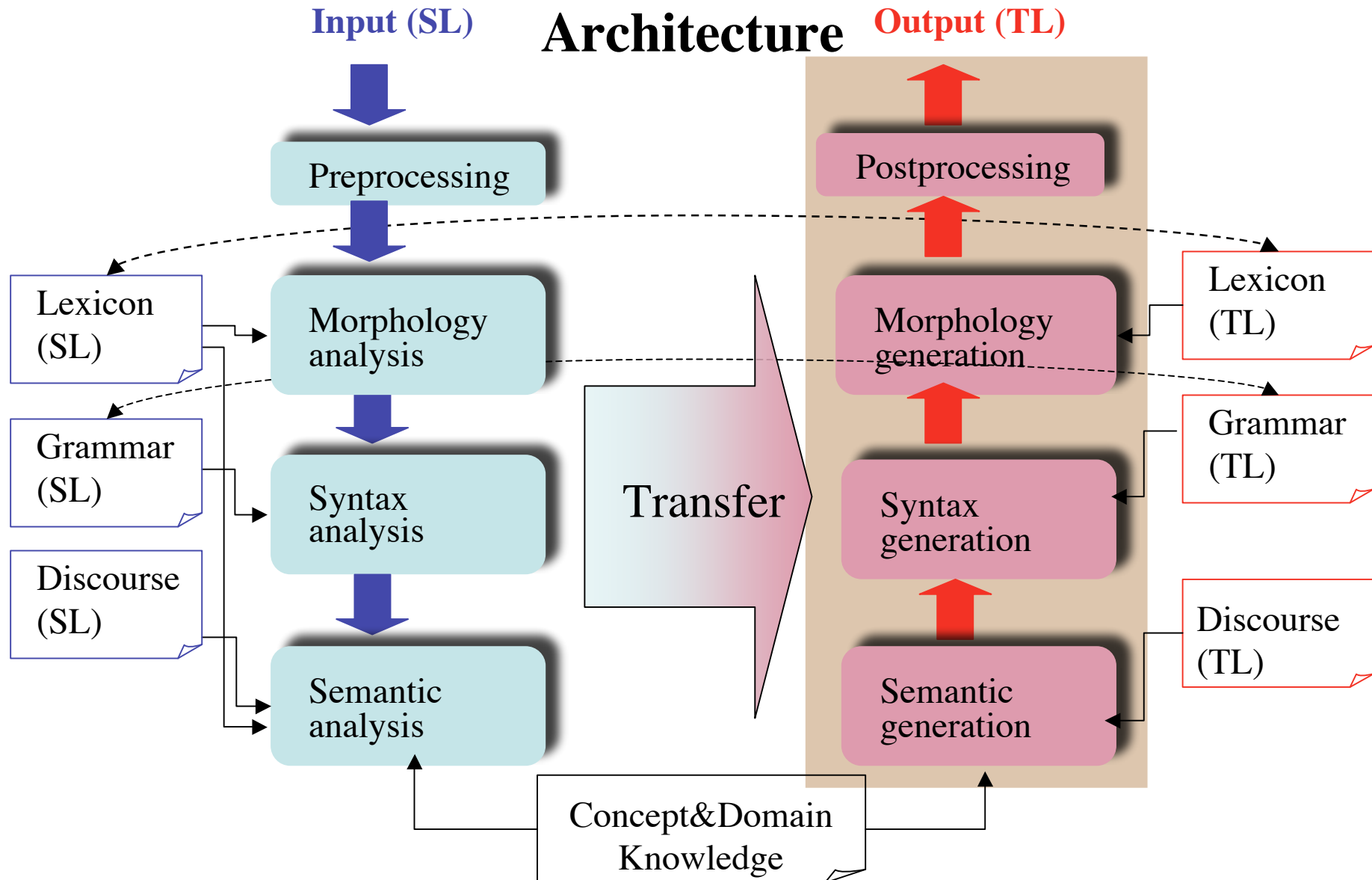
Semantic Transfer- Example of Head Switching

- Engl.: I think Carlos just fell - Sp.: Creo que Carlos acaba de caer
- QLF: En. [think,T,<speaker>,[and,[fall,E,carlos],[just,E]]].
- QLF: Sp. [creer,T,<speaker>,[acabar,A, carlos, [caer,E,carlos]]].
- Structurally the two QLFs are related as follows:
- [and,VPE,[just,E]] \Leftrightarrow [acabar,E,Subj,VPS] :- VPE \Leftrightarrow VPS

Structural Transfer with Interlingua

Tasks:

- Content analysis from the Lexicon, Morphology, Syntax, Semantics, Pragmatics
 - = Mapping of the Input on:
 - Presuppositions
 - Objects (e.g. = Variables),
 - Relationships (e.g. Roles),
 - Quantifiers (e.g.= Negation, Number)
 - Consistency check (e.g. Presupposition check)
 - Semantic extraction
 - Reordering of results in the generation phase
- 
- A series of red arrows originates from the first bullet point, "Content analysis from the Lexicon, Morphology, Syntax, Semantics, Pragmatics", and points to each of the four sub-bullets: "Presuppositions", "Objects (e.g. = Variables)", "Relationships (e.g. Roles)", and "Quantifiers (e.g.= Negation, Number)".



Generation in Direct Systems

- Is reduced to:
 - lexical substitution during the lexicon look-up
 - Local re-ordering
- Generation is based on SL as much as possible
- Nothing else is changed, unless it is strictly necessary for the production of an acceptable target language expression.

Generation in Transfer-based Systems -1-

- Split into two modules:
 - Syntactic generation
 - Morphological generation
- In **syntactic generation** the intermediate representation which is the output of analysis and transfer is converted into an ordered surface-surface-structure tree, with appropriate labelling of the leaves with target language grammatical functions and features
- Main task of syntactic generation is to order constituents in the correct sequence of the target language

Generation in Transfer-based Systems -2-

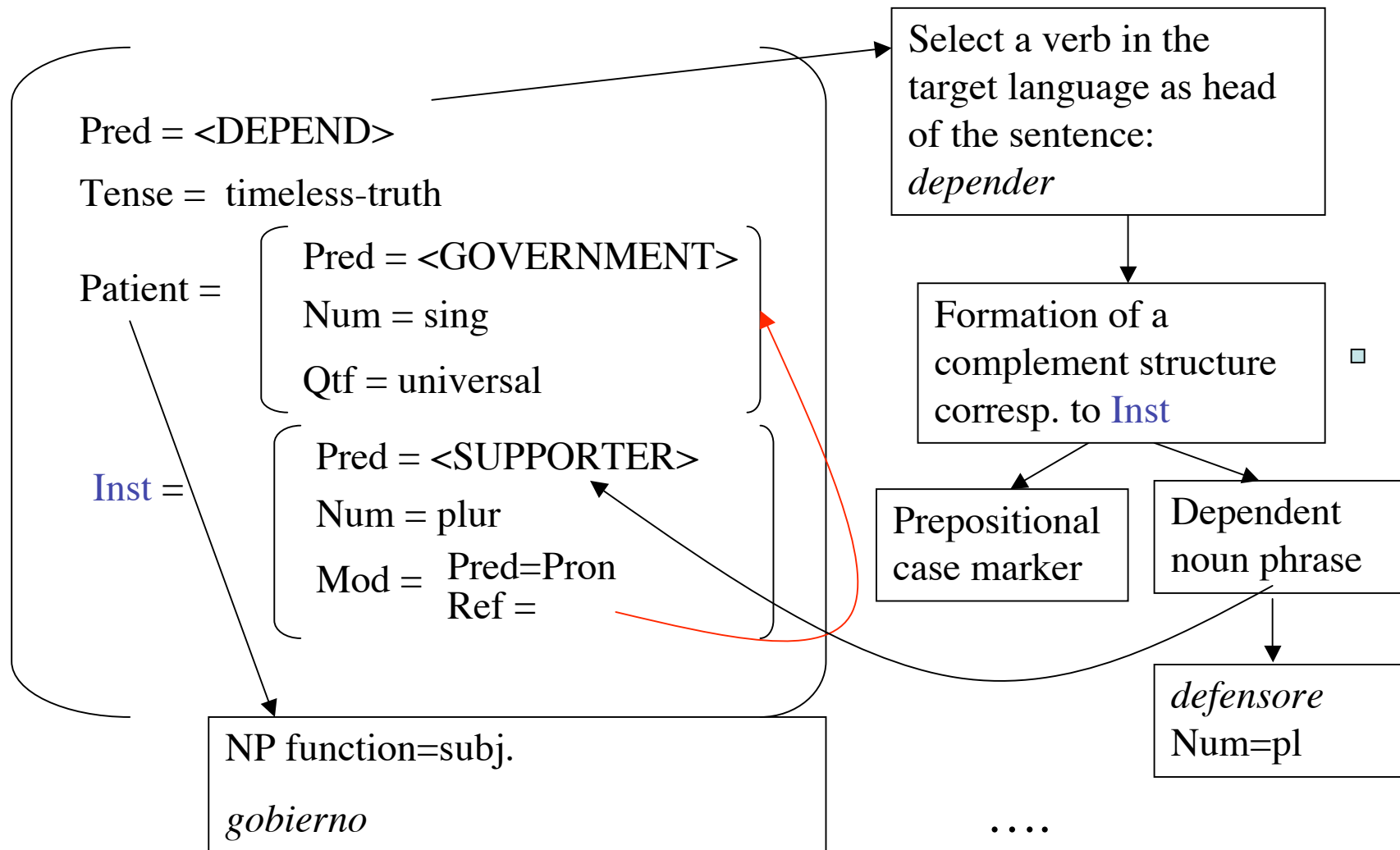
- E.g. For a sentence labelled as “passive” in the deep structure:
 - Syntactic generation creates a node for the auxiliary verb
 - Labelled with the appropriate tense information
 - Assign “past-participle” label to the main verb
- **Morphological generation** processes the resulting surface structure:
 - Interprets strings of labelled lexical items for target output
 - E.g: casa+pl = casas, ser+future+1stperson-sg. = seré
- Morphological generation can usually be handled by a combination of general and special-case procedures, on a word-by-word basis.

Generation in Interlingua-Systems

- Additionally to the syntactical and morphological generation, there is a **semantic generation** component
- The main task of the semantic generation is to find out which part of the interlingua expression should occur in the target sentence (e.g. not the existential presuppositions).
- The semantic generation produces as output a deep syntactic structure (i.e. a structure which has syntactic and semantic information, but is TL dependent)

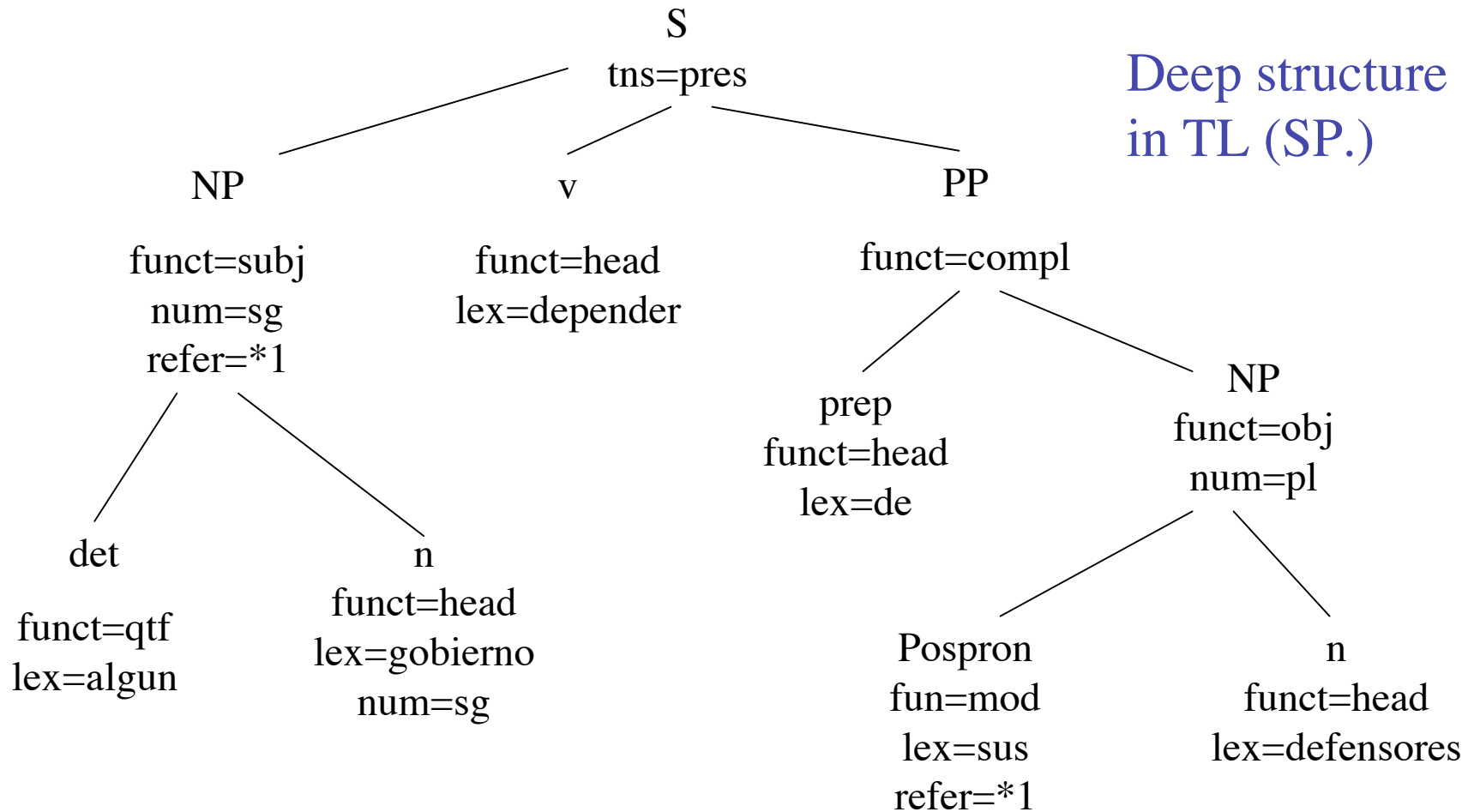
Generation in Interlingua Systems - Example

Each government depends on its supporters



Generation in Interlingua Systems -2-

Deep structure in TL (SP.)



Algun gobierno depende de sus defensores