Machine Translation
WiSe 2016/2017

HPI Hasso Plattner Institut

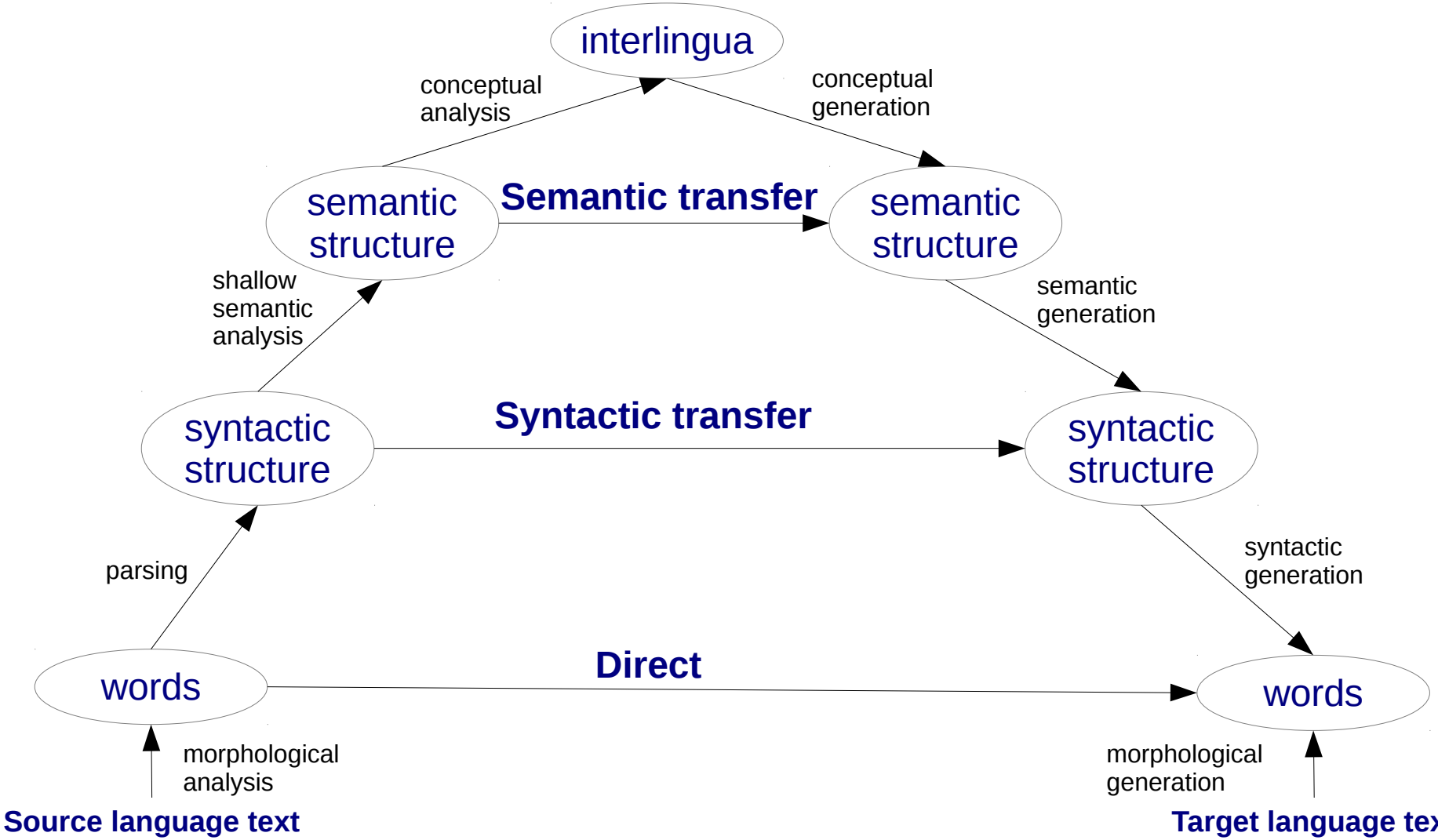
IT Systems Engineering | Universität Potsdam

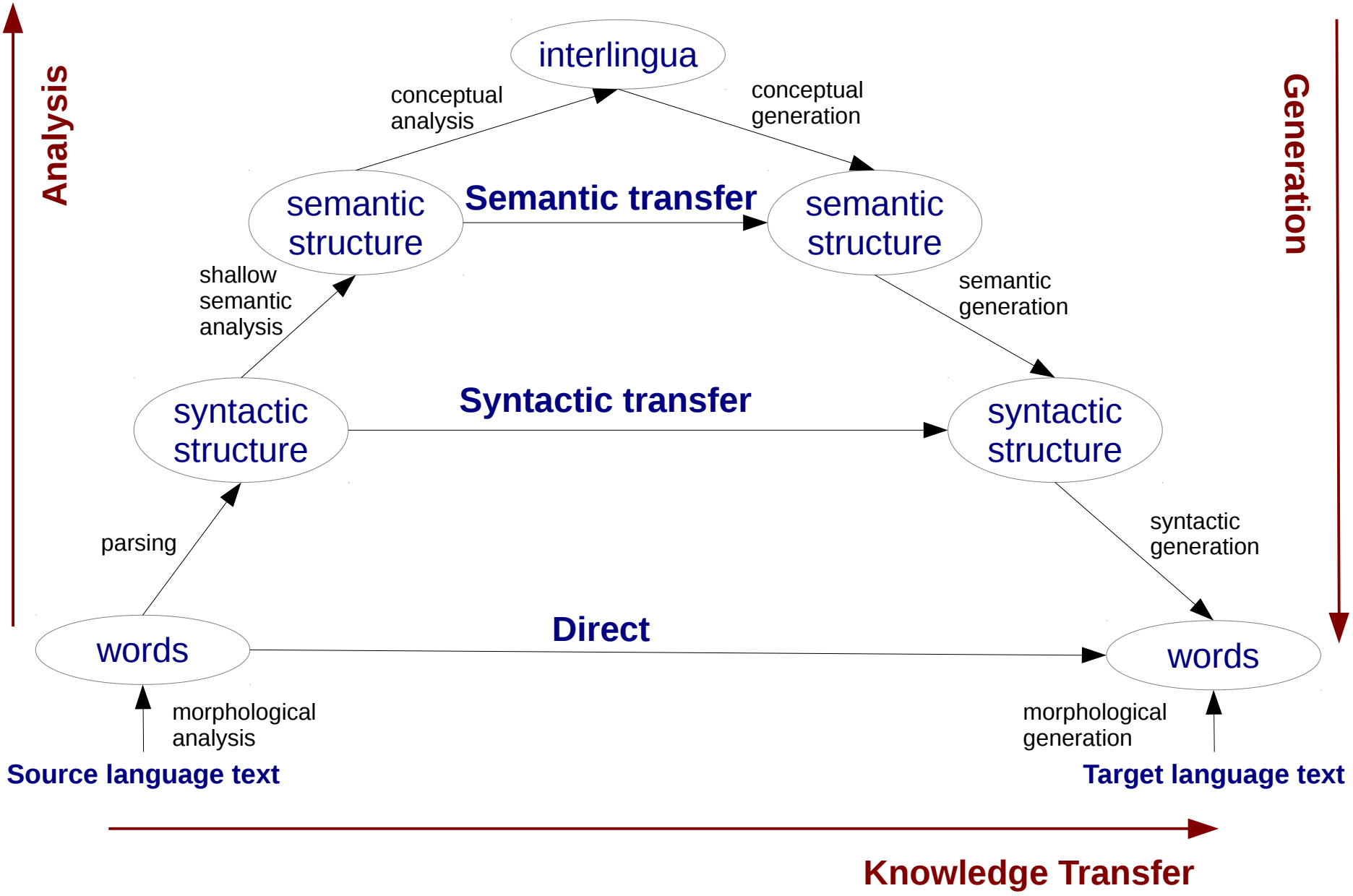
Rule-Based (Classical) Machine Translation

Dr. Mariana Neves

January 16th, 2017

Classical MT - Vauquois triangle




Analysis
Generation
interlingua
conceptual analysis
conceptual generation
semantic structure
Semantic transfer
semantic structure
shallow semantic analysis
semantic generation
syntactic structure
Syntactic transfer
syntactic structure
parsing
syntactic generation
words
Direct
words
morphological analysis
morphological generation
Source language text
Target language text
Knowledge Transfer

Classical MT

- Knowledge Transfer
  - Direct
    - Transfer knowledge for each word
  - Transfer
    - Transfer rules for parse trees or thematic roles
  - Interlingual
    - No specific transfer knowledge

Overview

- Direct
- Transfer
- Interlingual

Overview

- Direct
- Transfer
- Interlingual

Direct translation

- Word-by-word translation
- Transforming the source text into the target text

- Few linguistic analysis, no intermediate structure
  - shallow morphological analysis: verb tenses, stem

- The approach is not much used nowadays
- But it is the basis of many modern systems

Direct translation

- English to Portuguese:

Input: Mary didn't slap the green witch

Direct translation

- English to Portuguese:

Input:	Mary	didn't		slap	the green witch
After Morphology:	Mary	DO-PAST	not	slap	the green witch

Rules for past and negation (EN)

Direct translation

- English to Portuguese:

Input:	Mary	didn't	slap	the	green	witch
After Morphology:	Mary	DO-PAST	not	slap	the	green	witch
After lexical transfer:	Maria	PAST	não	dar uma tapa em	a	verde	bruxa

Comprehensive bilingual dictionary

Direct translation

- English to Portuguese:

Input:	Mary	didn't		slap	the	green	witch
After Morphology:	Mary	DO-PAST	not	slap	the	green	witch
After lexical transfer:	Maria	PAST	não	dar	uma tapa em	a	verde	bruxa
After local reordering:	Maria	não	dar	PAST	uma tapa em	a	bruxa	verde

Ordering rules (PT)

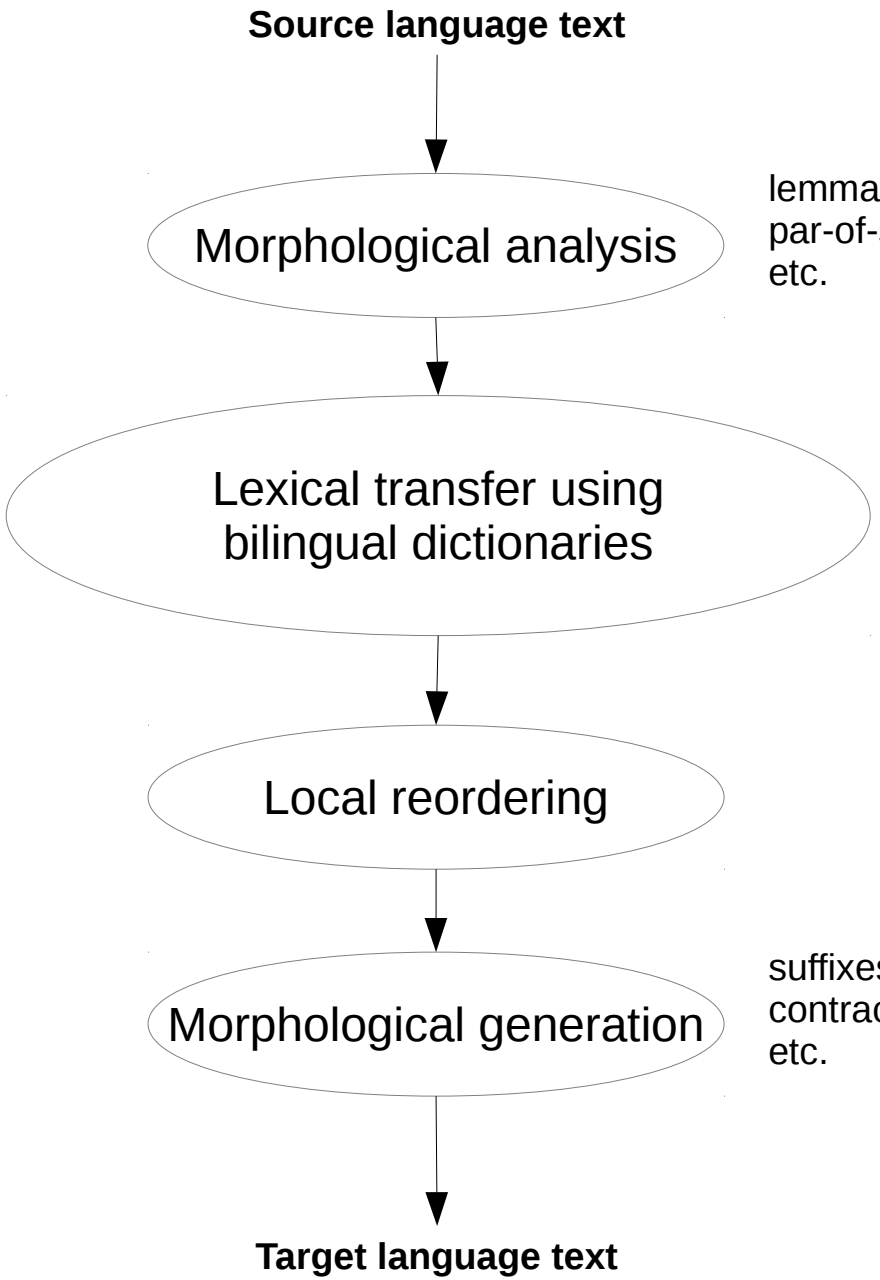
Direct translation

- English to Portuguese:

Input:	Mary	didn't	slap	the	green	witch	
After Morphology:	Mary	DO-PAST	not	slap	the	green	witch
After lexical transfer:	Maria	PAST	não	dar	uma	tapa	em	a	verde	bruxa
After local reordering:	Maria	não	dar	PAST	uma	tapa	em	a	bruxa	verde
After Morphology:	Maria	não	deu	uma	tapa	na	bruxa	verde		

Rules for past (irregular) and contractions (PT)

Direct Translation Workflow



Bilingual dictionaries - Wiktionary

Wiktionary
['wɪkʃənrɪ] *n.*,
a wiki-based Open Content dictionary

mit bloßem Auge (Deutsch)

- Chinesisch:
  - Mandarin (ISO 7098:1991): [1] 用肉眼 (yòng ròu yǎn) → zh
- Dänisch: [1] med det bare øje → da, med det blotte øje → da
- Englisch: [1] with the naked eye → en
- Französisch: [1] à l'œil nu → fr

Downloading Wiktionary [edit]

Q: Is it possible to download Wiktionary?

A: Yes. http://download.wikimedia.org/enwiktionary/ should have the latest copy of the main namespace. The cleanest navigation page is http://download.wikimedia.org/. Just download a *-articles.xml.bz2 file and some software to read it (for *nix, for Windows).

A: If you just want definitions, you can try http://tools.wmflabs.org/enwiktdefns/.

Q: Can I use data from Wiktionary in my program?

A: As long as you meet the conditions of the GNU Free Documentation License or Creative Commons Attribution/Share-Alike License, certainly.

(https://en.wiktionary.org/wiki/Wiktionary:Main_Page
https://de.wiktionary.org/wiki/mit_blo%C3%9Fem_Auge
https://en.wiktionary.org/wiki/Help:FAQ#Downloading_Wiktionary)

Bilingual dictionaries – Dict.cc

dict.cc
Deutsch-Englisch-Wörterbuch

Verben

to **slap** sb.	jdn. schlagen [ins Gesicht]	370
to **slap**	klatschen	238
to **slap**	patschen [ugs.]	44
to **slap**	anklatschen [ugs.] [ohne Sorgfalt ankleben]	18
to **slap**	batschen [südd.] [ugs.]	10
to **slap**	leicht schlagen	
to **slap**	mit der offenen Hand schlagen	
to **slap** sb.	jdm. eine langen [nordd.]	
to **slap** sb.	jdm. eine schallern [ugs.]	

(http://www.dict.cc/
http://www1.dict.cc/translation_file_request.php)

Bilingual dictionaries - BeoLingus

Aalstrich {m} [zool.] (dunkler RÃ¼ckenstreifen) :: dorsal stripe
Aalsuppe {f} [cook.] | Aalsuppen {pl} :: eel soup | eel soups
Aalterrine {f} [cook.] :: terrine of eel
Aapamoor {n} :: aapa mire; string bog
Aas {n}; Aasfleisch {n} [zool.] :: carrion
Aas {n} (Lederzurichtung) :: flesh; scrapings (leather dressing)
Aas fressen {vi} [zool.] :: to scavenge
Aaronsstab {m} (Zierleiste) [arch.] :: Aaron's rod
Aasfliege {f} [zool.] | Aasfliegen {pl} :: carrion fly; fleshfly; flesh fly | carrion flies; fleshflies; flesh flies
Aasfresser {m} [zool.] | Aasfresser {pl} :: scavenger; carrion eater; carrion feeder; scavenging animal | scavengers; carrion eaters; carrion feeders; scavenging animals

(http://dict.tu-chemnitz.de/
ftp://ftp.tu-chemnitz.de/pub/Local/urz/ding/de-en/)

Bilingual dictionaries - UWN/MENTA

- Towards a Universal Multilingual Wordnet
  - Words can be linked to Wordnet

(https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/uwn/
http://www.lexvo.org/uwn/)

Bilingual dictionaries - UWN/MENTA

WordNet
A lexical database for English

Verb

- S: (v) **slap** (hit with something flat, like a paddle or the open hand) *"The impatient teacher slapped the student"; "a gunshot slapped him on the forehead"*
  - *direct troponym* / *full troponym*
    - S: (v) cuff, whomp (hit with the hand)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (v) strike (deliver a sharp blow, as with the hand, fist, or weapon) *"The teacher struck the child"; "the opponent refused to strike"; "The boxer struck the attacker dead"*
  - *derivationally related form*
    - W: (n) slap [Related to: slap] (the act of smacking something; a blow delivered with an open hand)
    - W: (n) slap [Related to: slap] (a blow from a flat object (as an open hand))
    - W: (n) slapper [Related to: slap] (a hitter who slaps (usually another person) with an open hand) *"someone slapped me on the back and I turned to see who the slapper was"; "my father was the designated spanker in our family"*
  - ***sentence frame***
    - Somebody ----s something
    - Somebody ----s somebody
    - Something ----s something
    - [Applies to slap] The fighter managed to slap his opponent

(http://wordnetweb.princeton.edu/perl/webwn)

Rules

- Algorithm for translating „much“ and „many“ into Russian

function DIRECT_TRANSLATE_MUCH/MANY(word) **returns** Russian translation

if preceding word is *how* **return** *skol'ko*
else if preceding word is *as* **return** *stol'ko zhe*
else if word is *much*
  if preceding word is *very* **return** nil
  else if following word is a noun **return** *mnogo*
else /* word is many */
  if preceding word is a preposition and following word is a noun **return** *mnogii*
  else return *mnogo*

(Figure extracted from Jurafski and Martin 2009)

Rules

- Japanese counters

日本語	When to Use
人	To count the number of people
本	To count long, cylindrical objects such as bottles or chopsticks
枚	To count thin objects such as paper or shirts
冊	To count bound objects usually books
匹	To count small animals like cats or dogs
歳	To count the age of a living creatures such as people
個	To count small (often round) objects
回	To count number of times
ヶ所（箇所）	To count number of locations
つ	To count any generic object that has a rare or no counter

(http://www.guidetojapanese.org/learn/grammar/numbers)

Rules

- Japanese counters

	人	本	枚	冊	匹	歳	個	回	ヶ所（箇所）	つ
1	ひとり	いっぽん	いちまい	いっさつ	いっぴき	いっさい	いっこ	いっかい	いっかしょ	ひとつ
2	ふたり	にほん	にまい	にさつ	にひき	にさい	にこ	にかい	にかしょ	ふたつ
3	さんにん	さんぼん	さんまい	さんさつ	さんびき	さんさい	さんこ	さんかい	さんかしょ	みっつ
4	よにん	よんほん	よんまい	よんさつ	よんひき	よんさい	よんこ	よんかい	よんかしょ	よっつ
5	ごにん	ごほん	ごまい	ごさつ	ごひき	ごさい	ごこ	ごかい	ごかしょ	いつつ
6	ろくにん	ろっぽん	ろくまい	ろくさつ	ろっぴき	ろくさい	ろっこ	ろっかい	ろっかしょ	むっつ

(http://www.guidetojapanese.org/learn/grammar/numbers)

Rules

- Declesions in German
  - Article
  - Adjective (strong, mixed and weak inflections)

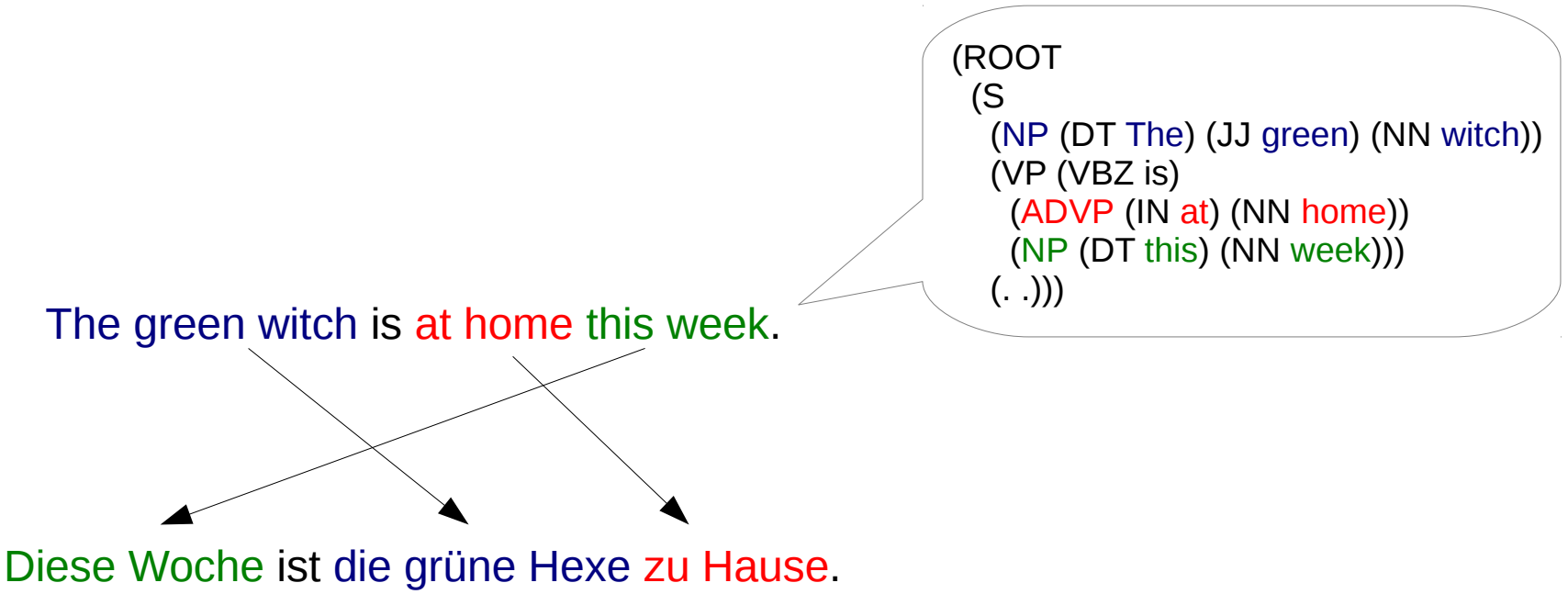
	Masculine	Neuter	Feminine	Plural
Nominative	alter	altes	alte	alten
Accusative	alten	altes	alte	alten
Dative	alten	alten	alten	alten
Genitive	alten	alten	alten	alten

(https://www.duolingo.com/skill/de/Adjectives:-Nominative-2)

Drawbacks of direct translation

- Need to rely on (language-specific)
  - comprehensive bilingual dictionaries
  - and morphological rules
- No parsing component and no knowledge on phrasing or grammatical structure
  - Cannot handle
    - long-distance reordering

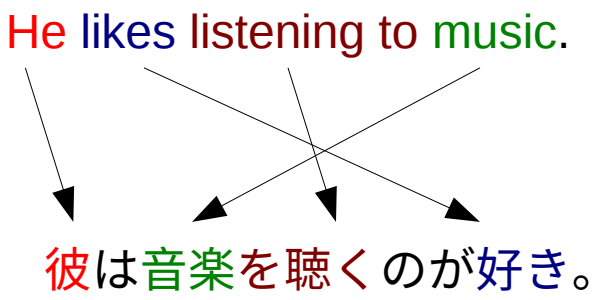
Drawbacks of direct translation - reordering



(http://nlp.stanford.edu:8080/parser/index.jsp)

Drawbacks of direct translation - reordering

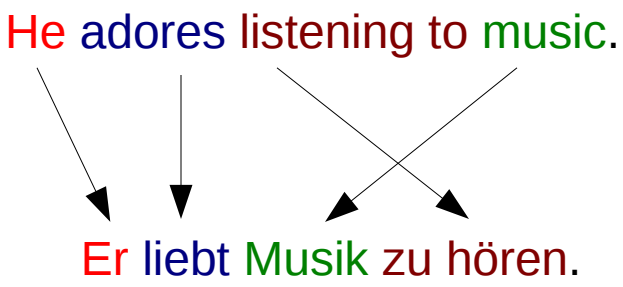
- Even more complex reordering
  - From SVO (Subject-Verb-Object)
  - To SOV (Subject-Object-Verb)



(http://translate.google.de)

Drawbacks of direct translation - reordering

- Even more complex reordering
  - From SVO (Subject-Verb-Object)
  - To SVO/SOV (Subject-Object-Verb)



(http://translate.google.de)

SOV, SVO, VSO, etc..

Word order	English equivalent	Proportion of languages		Example languages
SOV	"She him loves."	45%		Hindi, Latin, Japanese, Afrikaans
SVO	"She loves him."	42%		English, Hausa, Mandarin, Russian
VSO	"Loves she him."	9%		Biblical Hebrew, Irish, Filipino, Tuareg
VOS	"Loves him she."	3%		Malagasy, Baure
OVS	"Him loves she."	1%		Apalaí, Hixkaryana
OSV	"Him she loves."	0%		Warao

Frequency distribution of word order in languages surveyed by Russell S. Tomlin in 1980s.[1][2] (v · t · e)

(https://en.wikipedia.org/wiki/Subject%E2%80%93verb%E2%80%93object)

Overview

- Direct
- Transfer
- Interlingual

Transfer

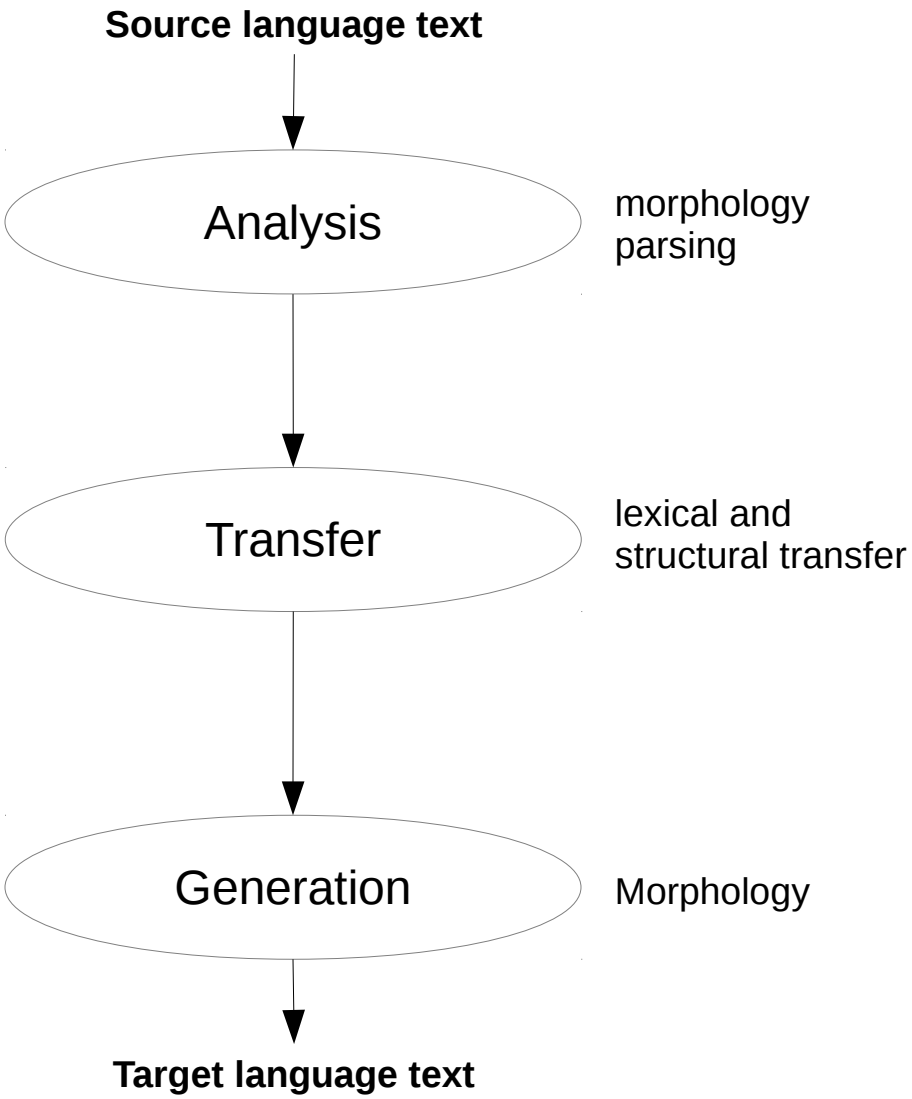
- It relies on the differences between the two languages
- Altering the structure of the source languages to make it conform to the rules of the target language

- Constractive knowledge
  - Knowledge about the difference between two languages

Transfer

- Transfer and Interlingual approaches use intermediate representations
  - Interlingual is language independent
  - Transfer depends on the language pair

Transfer Translation Workflow



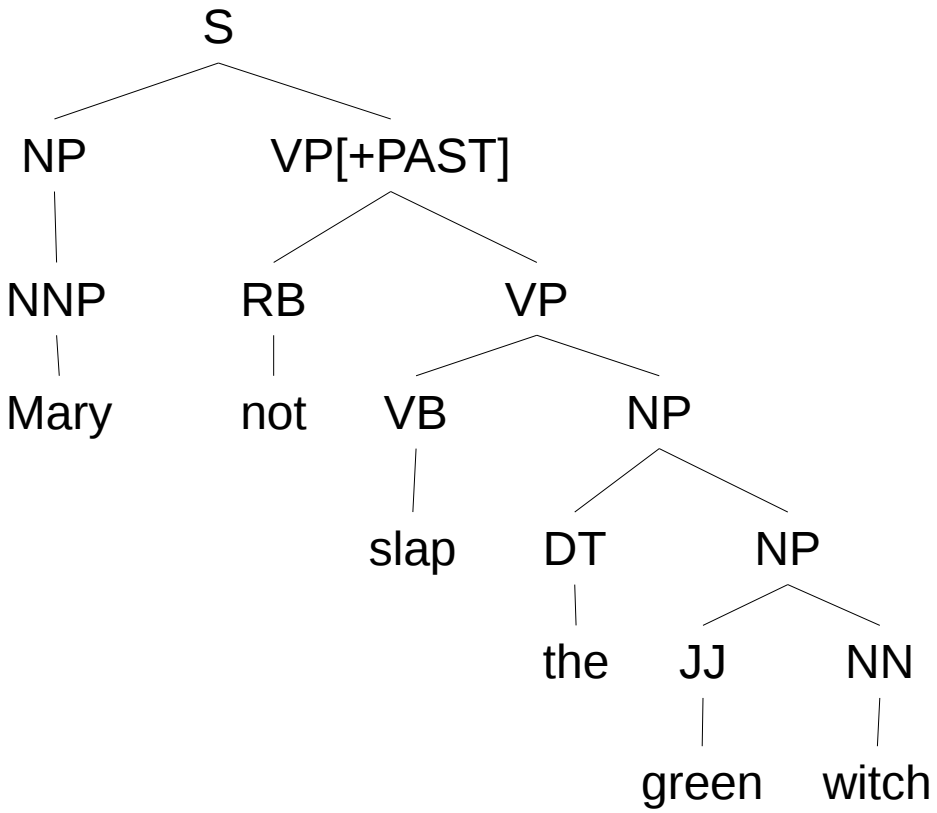
Morphology

- English: rules for past and negation

Input:	Mary	didn't		slap	the green witch
After Morphology:	Mary	DO-PAST	not	slap	the green witch

Analysis

- Parsing



Analysis

- A parsing should be adequate for machine translation purposes

```
(ROOT
  (S
    (NP (NNP John))
    (VP (VBD saw)
      (NP (DT the) (NN girl))
      (PP (IN with)
        (NP (DT the) (NNS binoculars))))
    (. .)))
```

John saw the girl with the binoculars.

John sah das Mädchen mit dem Fernglas.

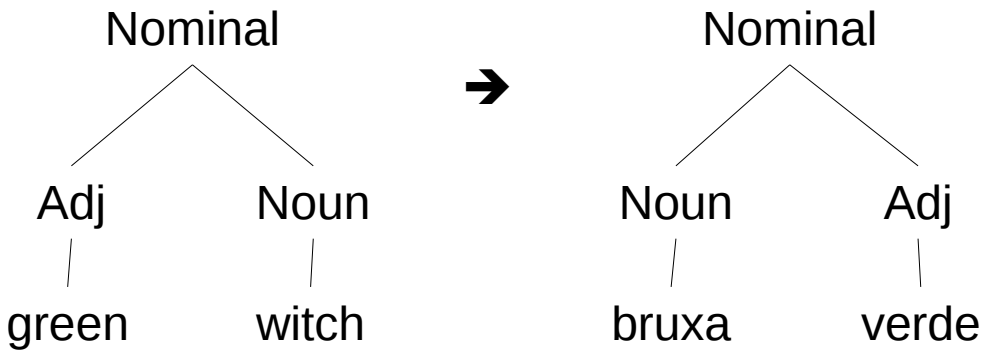
(http://nlp.stanford.edu:8080/parser/index.jsp http://translate.google.de)

Transfer

- Syntactic transfer
  - Modify the source parse tree to look like the target parse tree
    - Rules
- Lexical transfer
  - Modify the source words to the target words
    - Bilingual dictionary

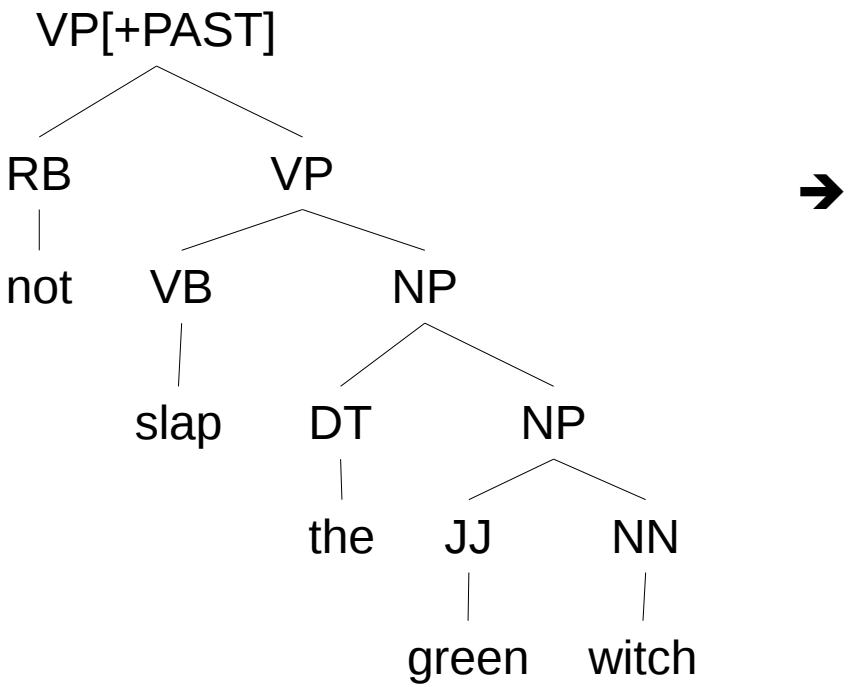
Rules for syntactic transfer

- Adjective-noun reordering



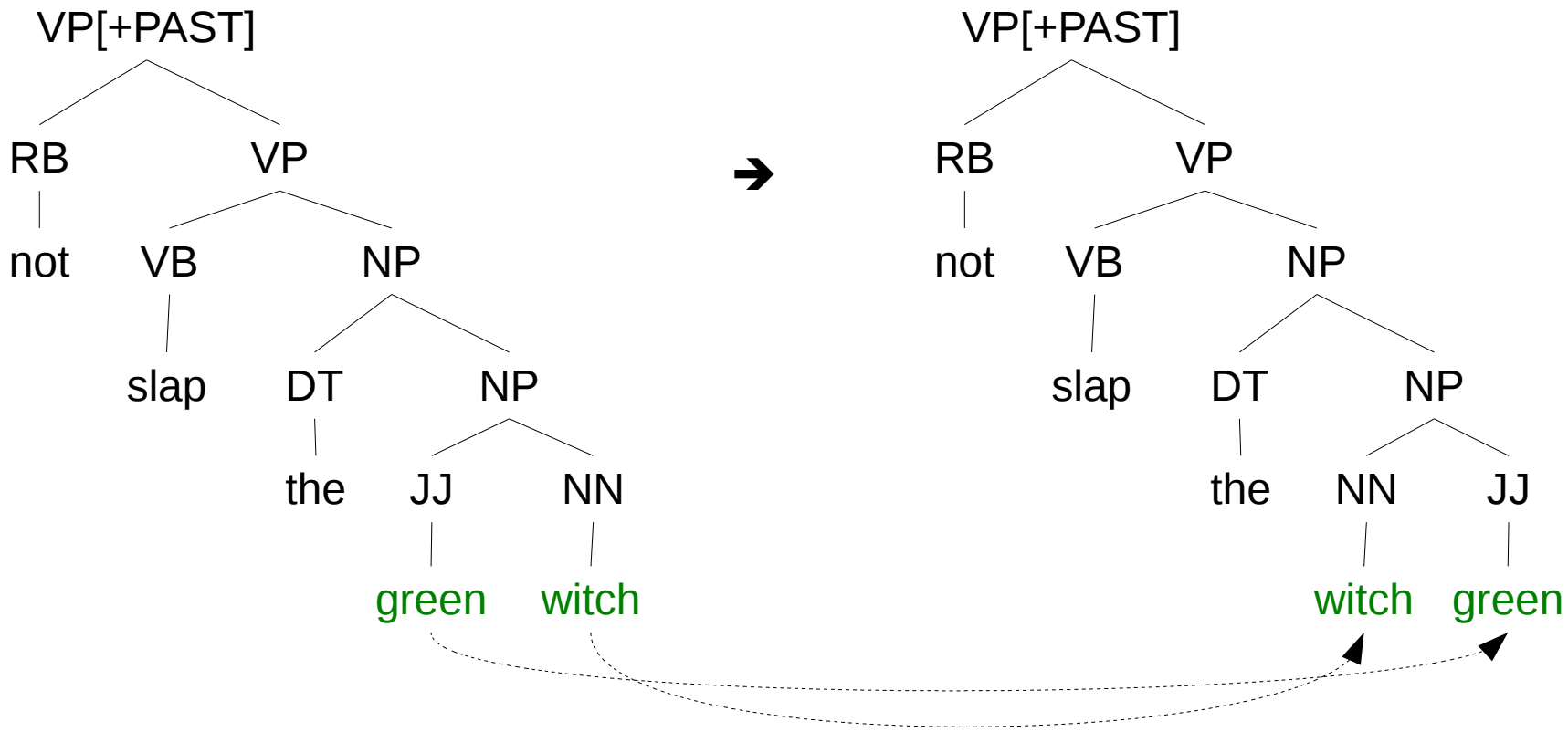
Syntactic transfer

- English to Portuguese



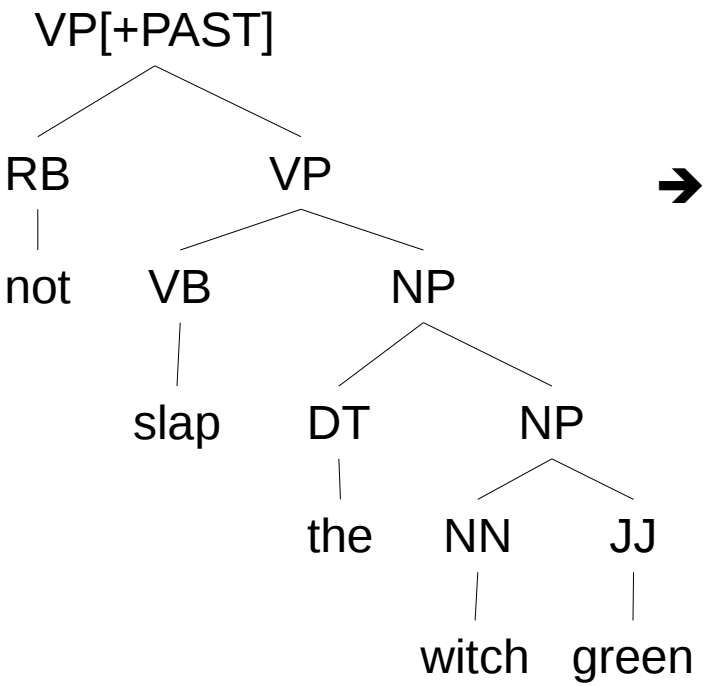
Syntactic transfer

- English to Portuguese: Adjective-noun reordering



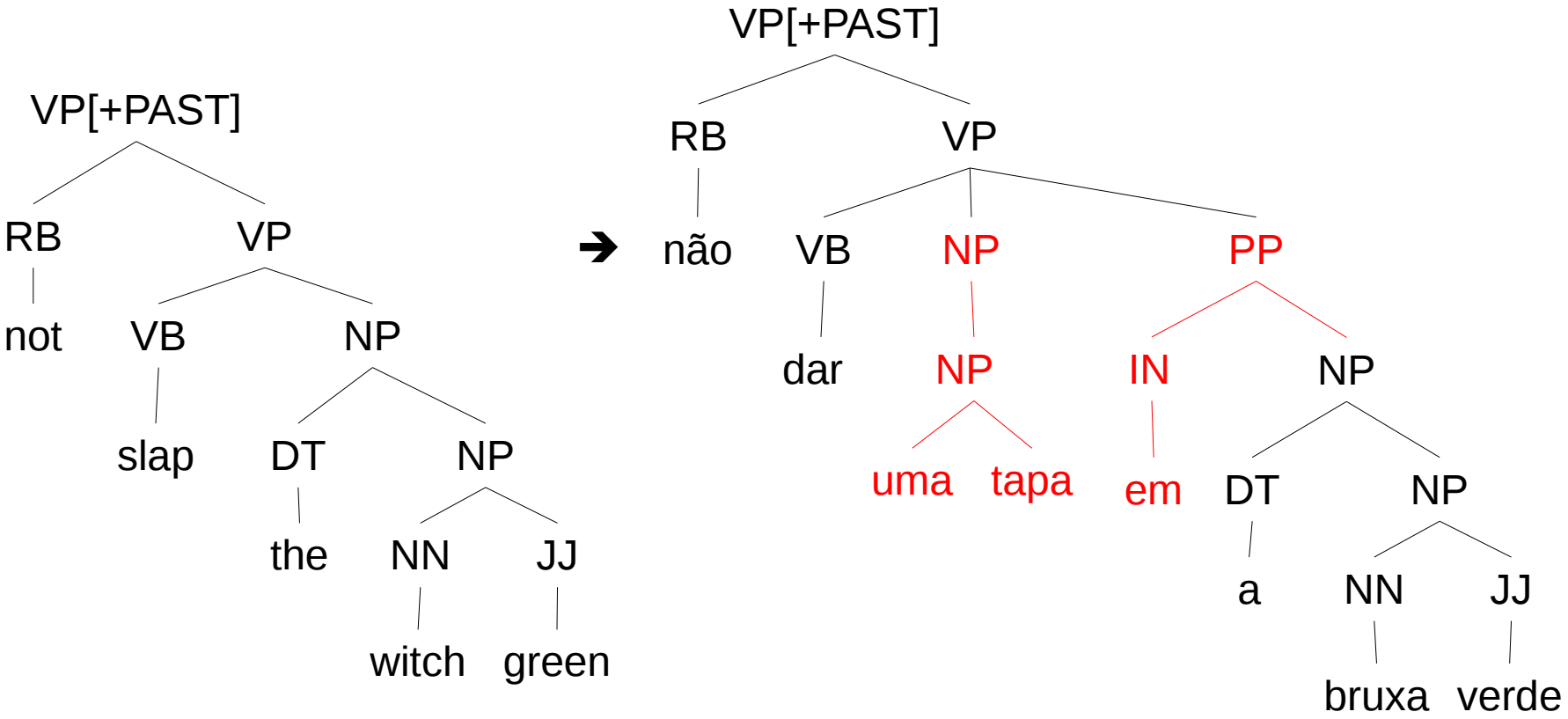
Lexical transfer

- English to Portuguese



Lexical transfer

- English to Portuguese



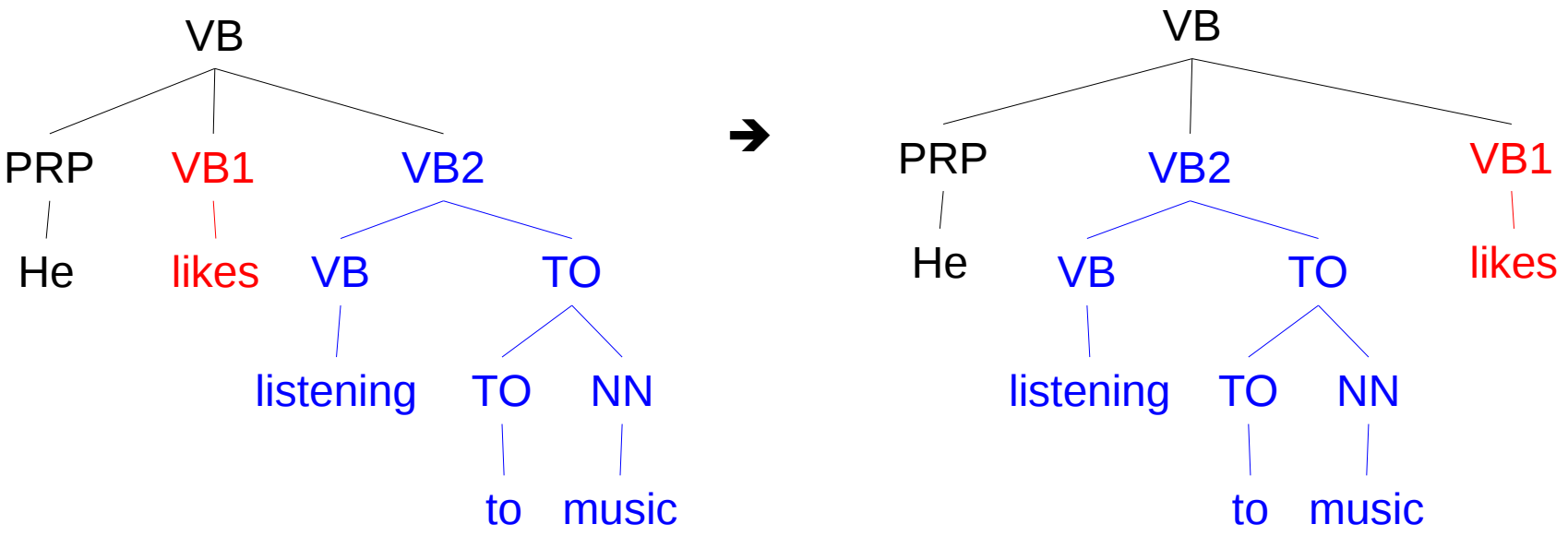
(http://lxcenter.di.fc.ul.pt/services/en/LXServicesParser.html)

Lexical transfer

- Lexical ambiguity
  - e.g., home
    - „nach Hause“ (going home)
    - „Heimfahrt“ (journey home)
    - „Heimat“ (home country)
    - „zu Hause“ (being at home)
- Idiomatic expressions
  - „dar uma tapa“ to „slap“

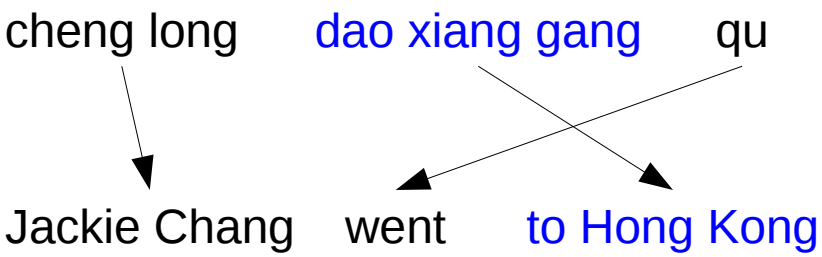
Syntactic Transfer

- Sometimes more complex rules are needed
- From English (SVO) to Japanese (SOV)



Syntactic Transfer

- Even more complex from English to Chinese
  - e.g., when describing goals



Syntactic Transfer

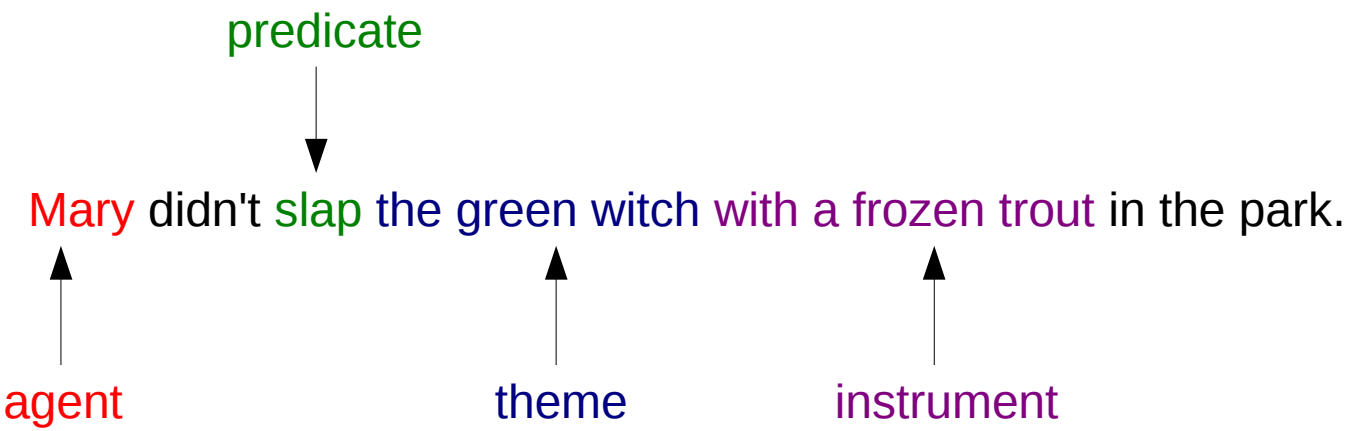
- Even more complex from English to Chinese
  - Various prepositional phrases
    - BENEFACTIVE PPs (before the verb)
    - DIRECTION and LOCATIVE PPs (before the verb)
    - RECIPIENT PPs (after the verb)
  - More understanding of the semantics

Semantic transfer

- Semantic role labeling (SRL)
  - Thematic role labeling
  - Case role assignment
  - Shallow semantic parsing

Semantic role labeling

- Determining which constituents (phrases) are semantic arguments for a given predicate (verb)
- Determining the appropriate role for each of the arguments



Resources for SRL – PropBank/VerbNet

- Around 5,000 verb senses
- „slap“ verb:
  - Roleset id: **slap.01**
  - Role:
    - **Arg0-PAG**: agent, hitter - animate only!
    - **Arg1-PPT**: thing hit
    - **Arg2-MNR**: instrument, thing hit by or with

(http://verbs.colorado.edu/verb-index/index.php
http://verbs.colorado.edu/propbank/framesets-english/slap-v.html)

Syntactic analysis

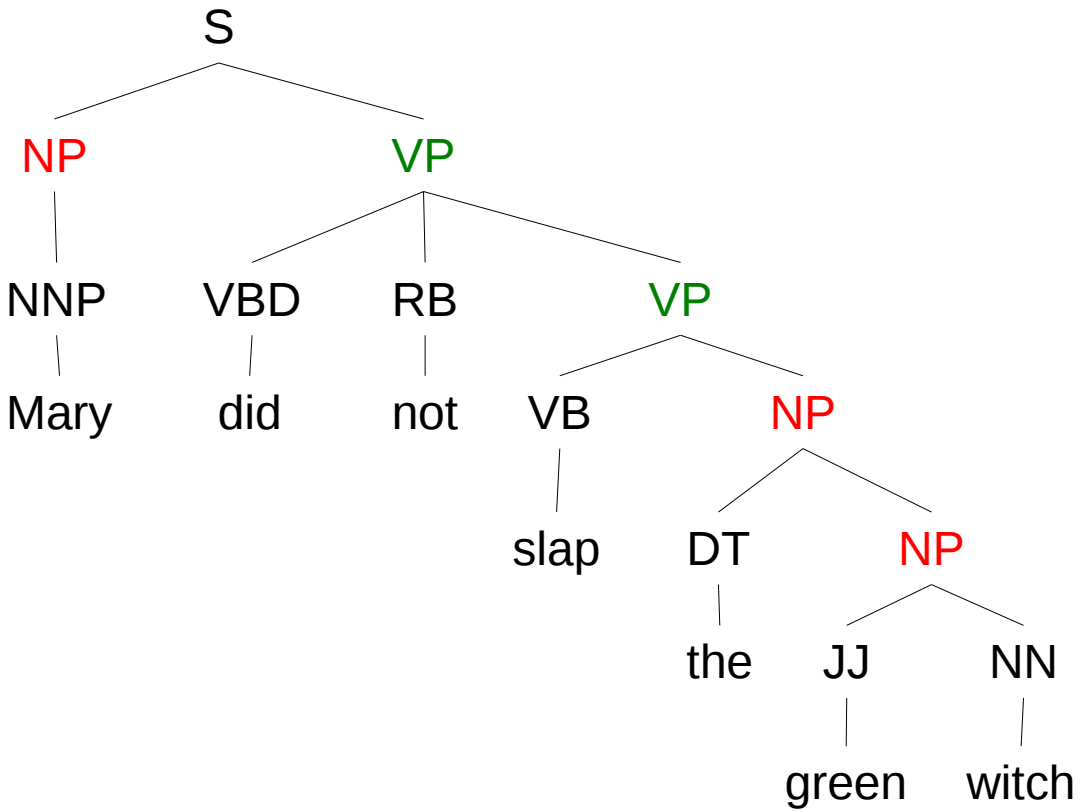
- SRL needs to rely on syntactic analysis
  - Chunking or shallow parsing

Mary/NP did/VP not/VP slap/VP the/NP green/NP witch/NP.

(http://nlp.stanford.edu:8080/parser/index.jsp)

Syntactic analysis

- SRL needs to rely on syntactic analysis
  - Parsing tree



(http://nlp.stanford.edu:8080/parser/index.jsp)

Syntactic analysis

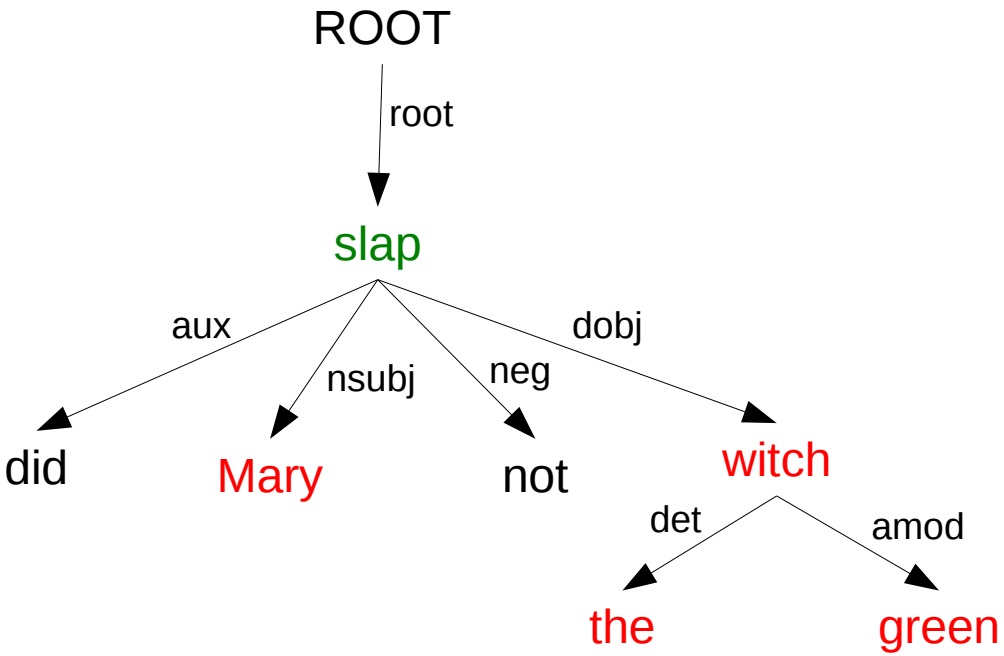
- SRL needs to rely on syntactic analysis
  - Parsing tree

```
(ROOT
  (S
    (NP (NNP Mary))
    (VP (VBD did) (RB n't)
      (VP (VB slap)
        (NP (DT the) (JJ green) (NN witch))
        (PP (IN with)
          (NP
            (NP (DT a) (JJ frozen) (NNS trout))
            (PP (IN in)
              (NP (DT the) (NN park)))))))
    (. .)))
```

(http://nlp.stanford.edu:8080/parser/index.jsp)

Syntactic analysis

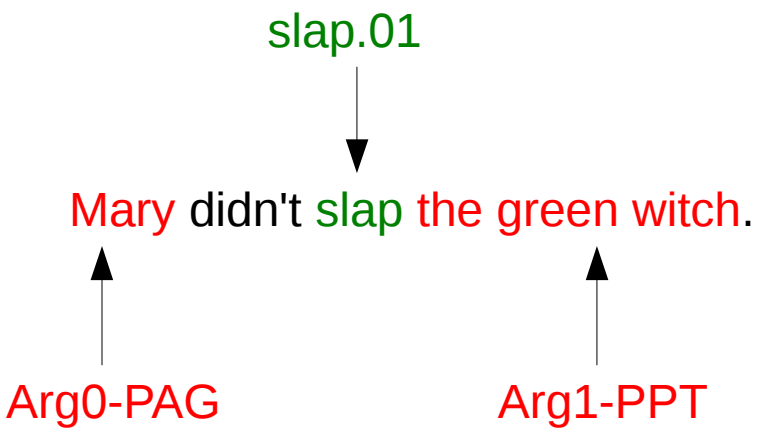
- SRL needs to rely on syntactic analysis
  - dependency tree



(http://nlp.stanford.edu:8080/parser/index.jsp)

Semantic role labeling

- Finding to the „slap“ predicate and respective arguments



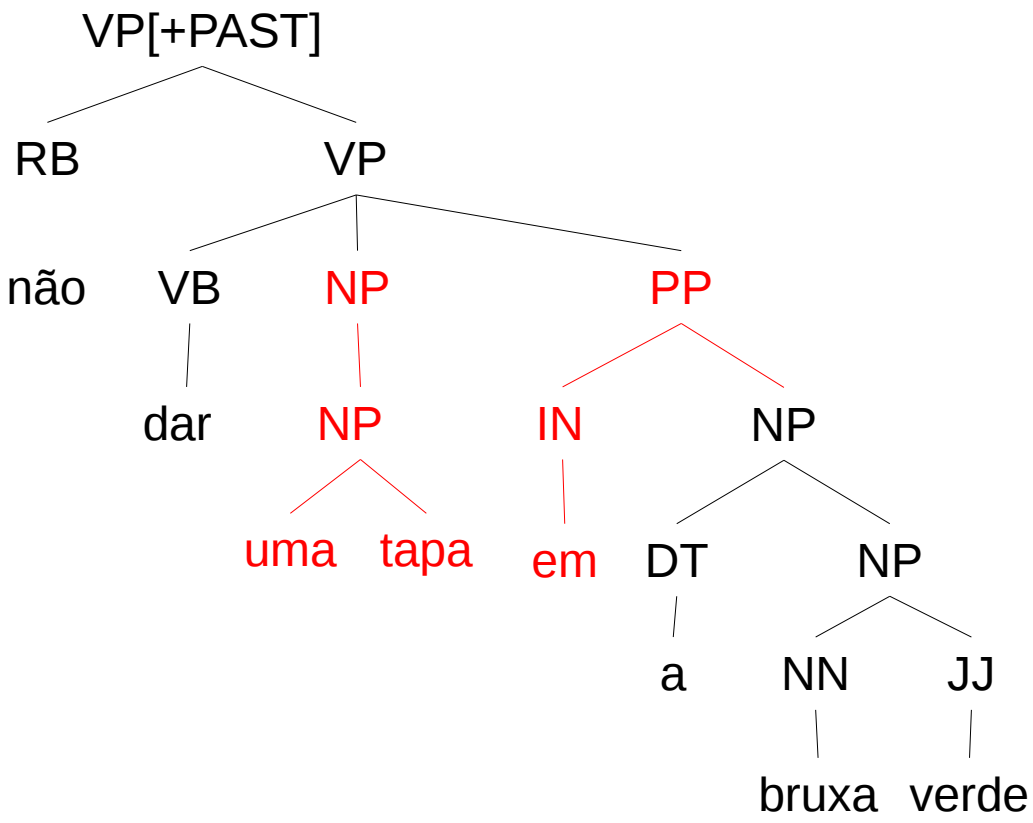
Semantic role labeling

- Methods are usually based on supervised machine learning and annotated corpora are necessary

- It has the potential to improve performance in any language-understanding task
  - Question answering
  - Information extraction

Generation

- After syntactic and lexical transfer



(http://lxcenter.di.fc.ul.pt/services/en/LXServicesParser.html)

Generation

- Morphology: rules for past and contraction

Input:	Maria não dar PAST uma tapa em a bruxa verde
After Morphology:	Maria não deu uma tapa na bruxa verde

Drawbacks of transfer translation

- It needs complex rules to transform the parse trees from one language to the other
  - Not only reordering
  - But also adding branches to the tree („uma tapa em“)

Hybrid approach: Direct+Transfer

- It is used by many commercial systems
  - e.g., Systran system

- Workflow
  - Shallow parsing
    - Morphological analysis
    - Part-of-speech tagging (nouns, verbs, adjectives, etc.)
    - Chunking (noun phrases, verbal phrases, etc.)
    - Shallow dependent parsing (subjects, passives, etc.)

Hybrid approach: Direct+Transfer

- Workflow (continuation)
  - Transfer
    - Translation of idioms
    - Word sense disambiguation
    - Assignment of prepositions according to verbs
  - Synthesis/Generation
    - Lexical translation (rich bilingual dictionary)
    - Reorderings
    - Morphological generation

Overview

- Direct
- Transfer
- Interlingual

Interlingual

- Shortcoming of transfer
  - Lexical and syntactic transfer for each pair of rules
  - Not feasible for multilingual environments:
    - European Union
    - Translation of manuals
    - Web pages

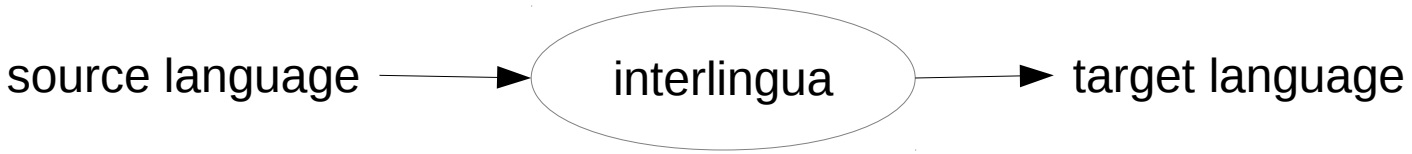
Interlingual

- No direct transformation of source language to target language
- Two steps transformation
  - Extract the meaning from the source language
  - Express the meaning in the target language

- Rely only on analysis and generation tools for each language
- Amount of knowledge is proportional to the number of languages, rather the square of it

Interlingua

- It is the representation of the meaning
- It is a language-independent canonical form
- It represents all sentences with the same meaning in the same way



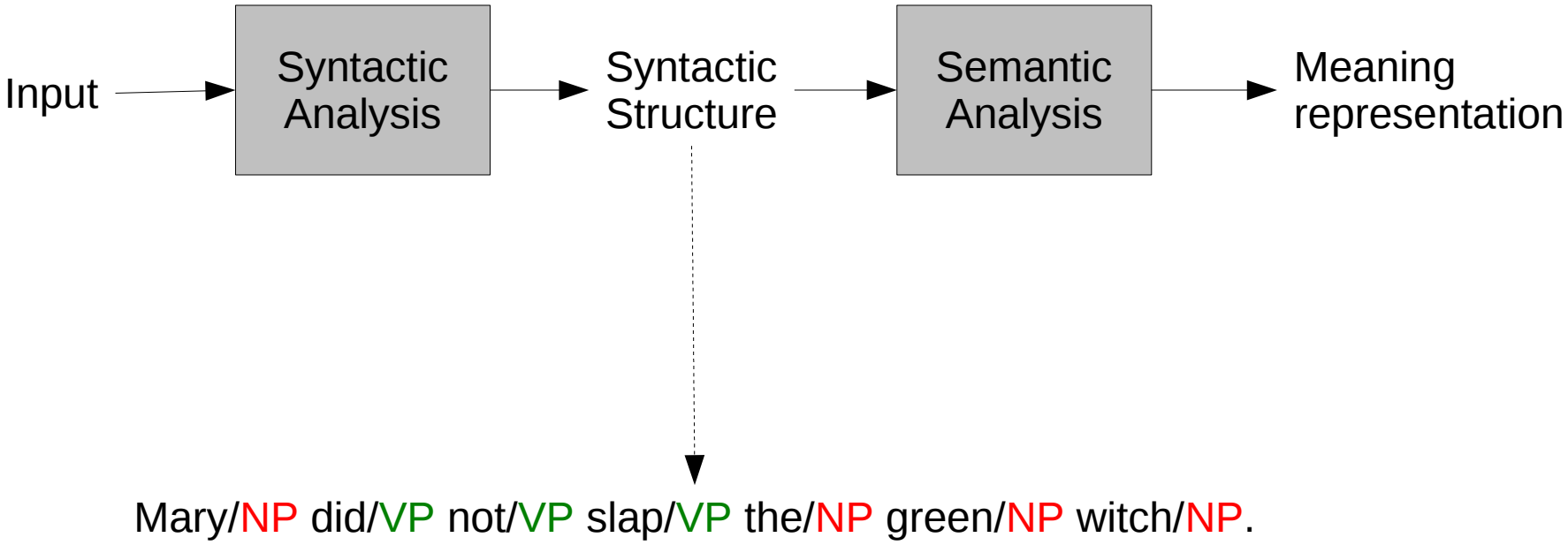
Semantic analyzer

- Principle of the compositionality
  - The meaning of a sentence can be constructed from the meaning of its parts
  - Ordering and grouping of words
  - Relations among the words

} Syntactic analysis

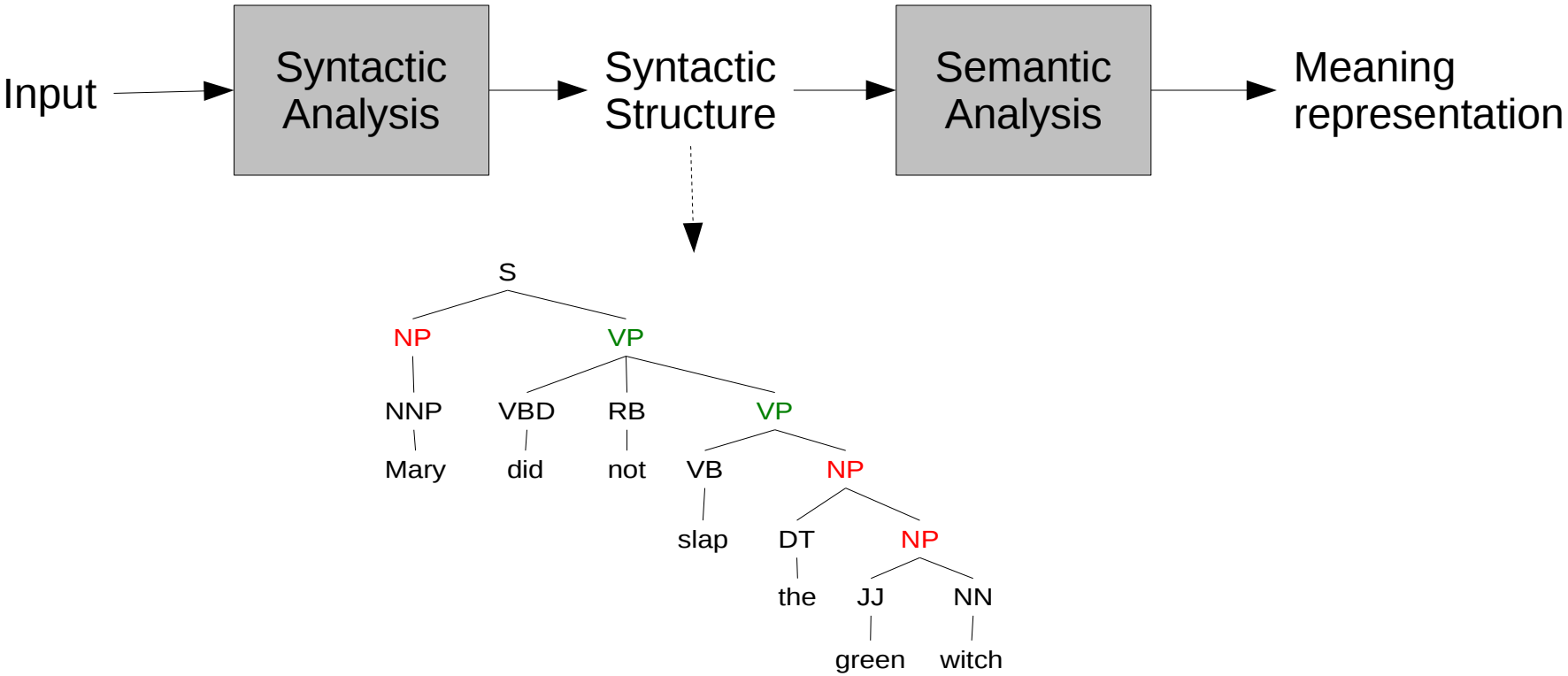
Workflow of a semantic analyzer

- Based on chunking



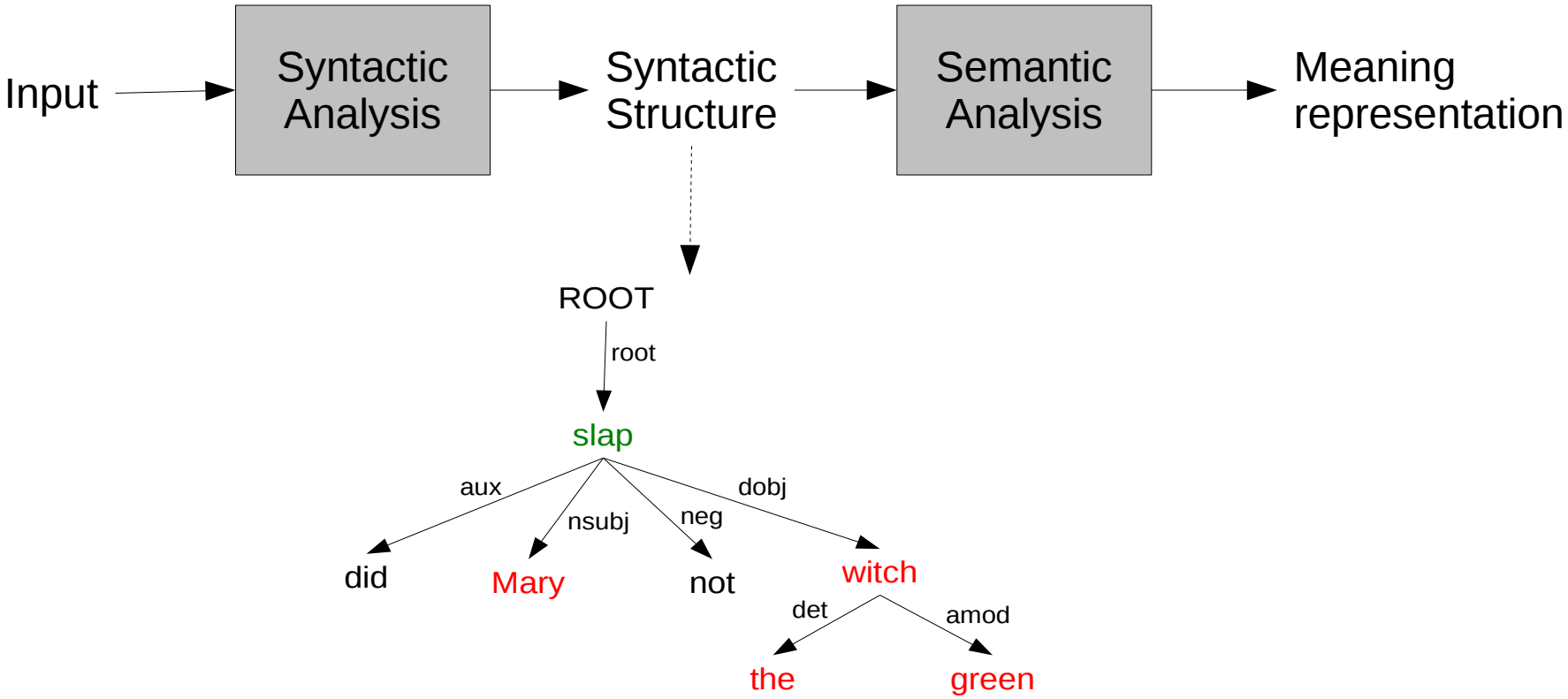
Workflow of a semantic analyzer

- Based on parsing tree



Workflow of a semantic analyzer

- Based on dependency tree



Dealing with ambiguities

- Syntactic ambiguity
  - „The woman saw the man with the telescope.“
- Lexical ambiguities
  - „fall“ (season, verb)
- Anaphoric ambiguities
  - „Alice understands that you like your mother, but **she** …“

- We assume here that all these ambiguities have been solved!

Interlingua - event-based representation

- Event
  - Aspectual: slap, not a kick
  - Negation: it is negated
  - Temporal: in the past, but details not specified
  - Entities: Mary, the witch

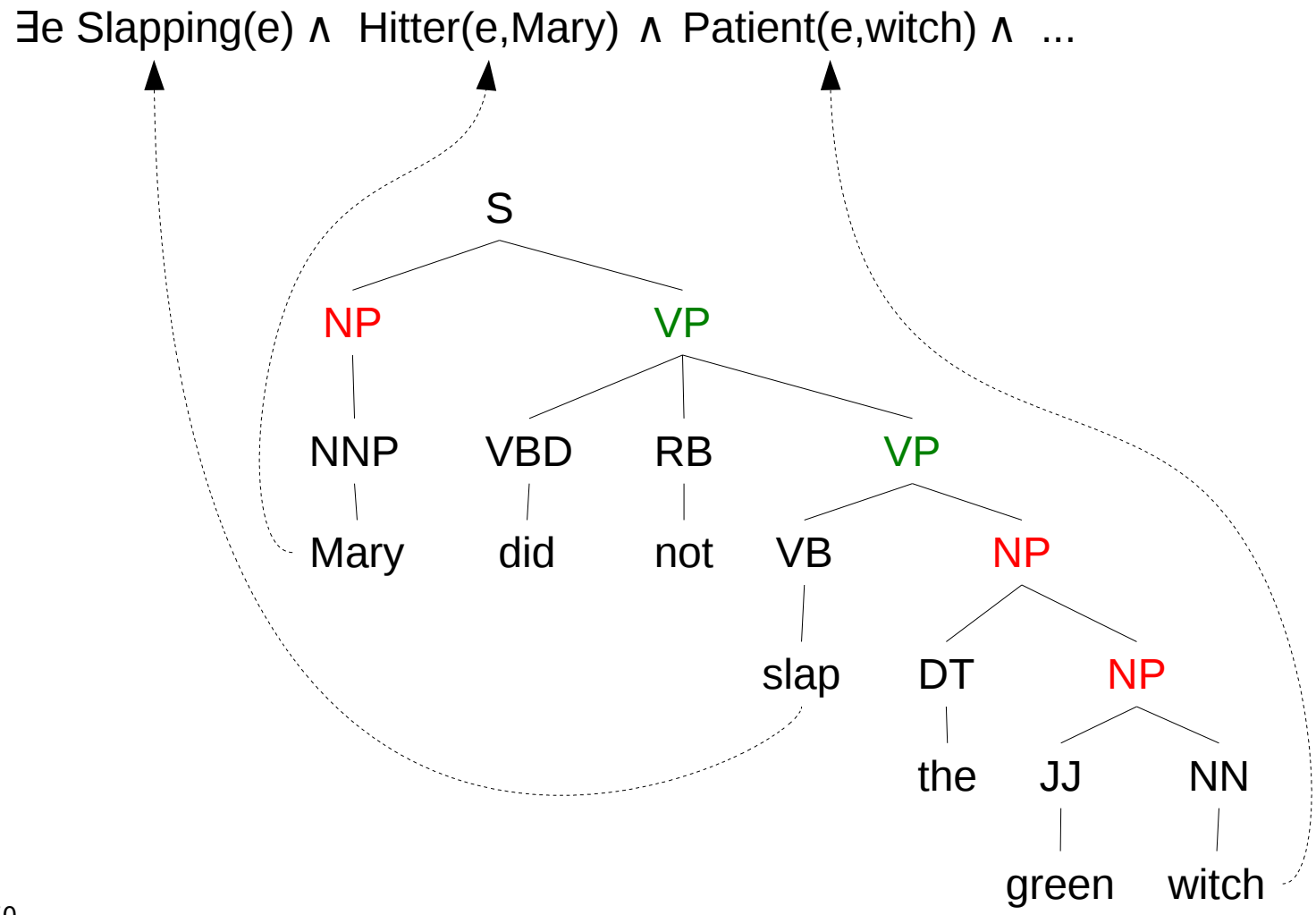
- Relations between the entities
  - has-color: green witch

Event-based representation

- Events are linked to their arguments through a small fixed set of thematic roles

EVENT	SLAPPING
AGENT	MARY
TENSE	PAST
POLARITY	NEGATIVE
THEME	WITCH
	DEFITIVENESS DEF
	ATTRIBUTES [**HAS-COLOR** GREEN]

Interlingua – logical notation



Interlingua – logical notation

- First-order logic (FOL)
  - First-order predicate calculus
  - Lower predicate calculus
  - Quantification theory
  - Predicate logic

First-order logic

A big boxer dates Mia in the park on Sunday.

$$\exists e(date(e) \wedge \exists b(boxer(b) \wedge big(b) \wedge subj(b, e)) \wedge$$

$$obj(Mia, e) \wedge place(park, e) \wedge time(Sunday, e))$$

(http://kilian.evang.name/publications/bathesis.pdf)

First-order logic

Mary did not slap the green witch.

$\nexists e(slap(e) \wedge subj(Mary, e) \wedge \exists b(witch(b) \wedge green(b)) \wedge obj(b, e) \wedge time(past, e))$

Semantic role labeling

- Based on predicates (verbs) and roles (phrases)

EVENT	SLAPPING
AGENT	MARY
TENSE	PAST
POLARITY	NEGATIVE
THEME	WITCH
	DEFITIVENESS DEF
	ATTRIBUTES [**HAS-COLOR** GREEN]

Semantic role labeling

- Based on predicates (verbs) and roles (noun phrases)

EVENT	SLAPPING (slap.01)
AGENT	MARY (Arg0-PAG)
TENSE	PAST
POLARITY	NEGATIVE
THEME	WITCH (Arg1-PPT)
	DEFITIVENESS DEF
	ATTRIBUTES **HAS-COLOR** GREEN

Generation

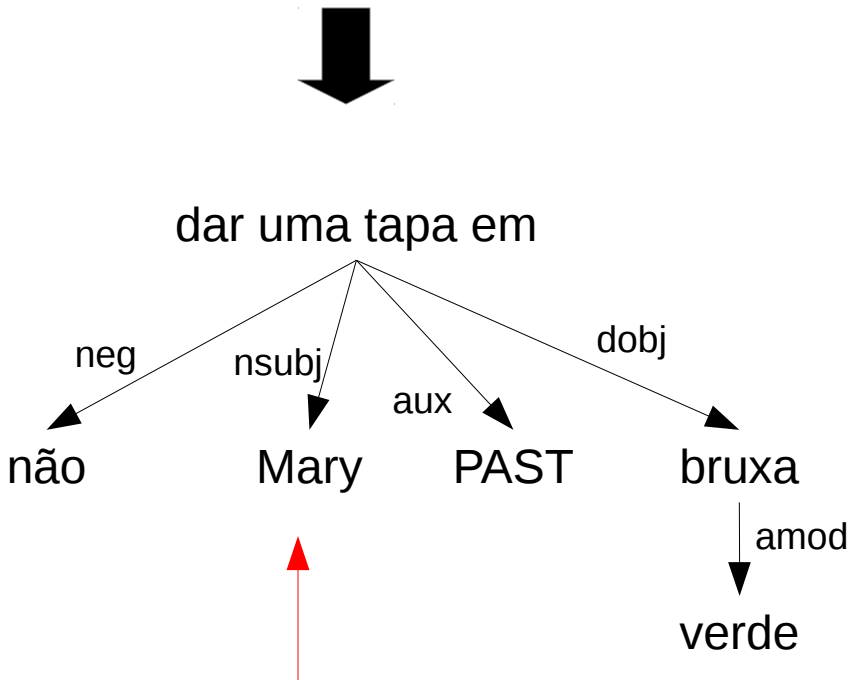
- No lexical and syntactic transfer rules are necessary

- Language generation using general natural language processing techniques and tools

Generation from logical notation

- Conversion to a dependency tree

$\nexists e(slap(e) \wedge subj(Mary, e) \wedge \exists b(witch(b) \wedge green(b)) \wedge obj(b, e) \wedge time(past, e))$



Generation - morphology

- Portuguese: rules for past and negation

Input:	Mary não PAST dar uma tapa em bruxa verde
After Morphology:	Mary não deu uma tapa na bruxa verde

Generation from event-based representation

- Based on SVO grammar rules

Generation - morphology

- Portuguese: rules for past and negation

Input:	Mary	não	PAST dar	uma tapa	em a bruxa verde
After Morphology:	Mary	não	deu	uma tapa	na bruxa verde

Drawbacks of interlingual

- Exhaustive analysis of the semantics of the domain
- Formalization into an ontology

- Only feasible for limited domains
  - Air travel, weather reports, hotel reservation, etc.
- Or for domains where such an ontology is available
- Or when using controlled languages

Drawbacks of interlingual

- Interlingua needs to represent the many verb senses:

	English	German
HAVE-A-PROPOSITION-IN-MEMORY	to know	wissen
BE-ACQUAINTED-WITH-ENTITY	to know	kennen

Drawbacks of interlingual

- Unnecessary disambiguation across many languages
  - Chinese has concepts for ELDER_BROTHER and YOUNGER_BROTHER

Summary

- Direct
  - It uses little analysis (morphology, part-of-speech tagging)
  - It uses lots of knowledge transfer
- Transfer
  - It relies on syntactic parsing, semantic role labeling
  - It needs on lexical and syntactic transfer for each language pair
- Interlingual
  - It uses a representation of the meaning (interlingual)
  - It relies only on language-specific NL processing and generation tools

Suggested reading

- Speech and Language Processing (chapter 25.2)
  - Daniel Jurafsky and James H. Martin