Basic Concepts and Technologies of 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

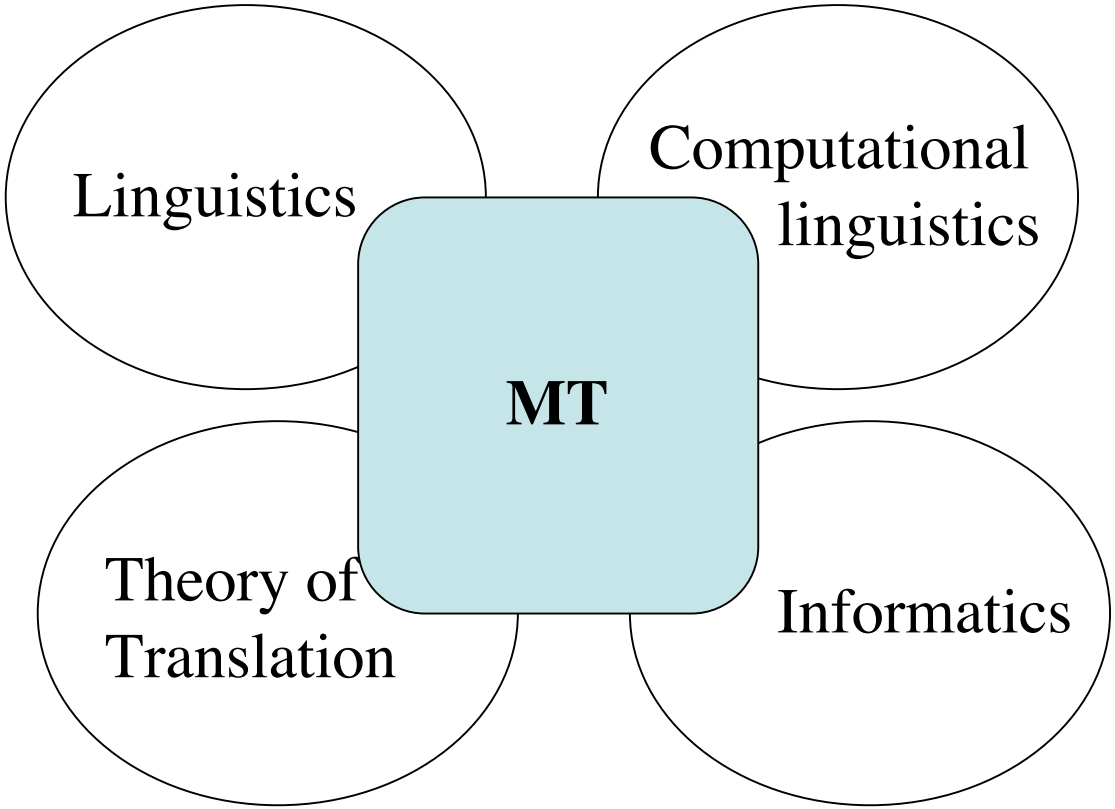
Some slides I owe with explicit permission from Walther v. Hahn

Abreviations

- MT = Machine Translation
- MAT = Machine Aided Translation
- MAHT = Machine Aided Human Translation
- HAMT = Human aided Machine Translation
- SL, TL = Source Language / Target Language

Machine Translation as a Discipline

MT is not a discipline by itself, but an application of several disciplines



Why do we need MT?

- Worldwide the translation market has a value (in million $) von
  1989 20
  1990 500
  2003 2000 The approximate growth is 20% each year
- Already in 1986 worldwide more than 500 Mio pages of translations, more as 100 Mio in Europe.

  1% > “nice Literature"

  30% official (state) places

  50% Industry and commerce (mostly technical documentation)

- Economy of time by using MT-system Systran: 75%
- Improvement of services by using MAT-systems (following German Airbus): 20%

- Systran translated in 1994 140 000 pages EU-Documents. 80 % of the EU-Documents between Spanish and French are result of machine translation
- We cannot educate so many translators as we need
- The situation became more critical with the enlargement of the EU: for the moment there are 20 official languages and the number will increase!

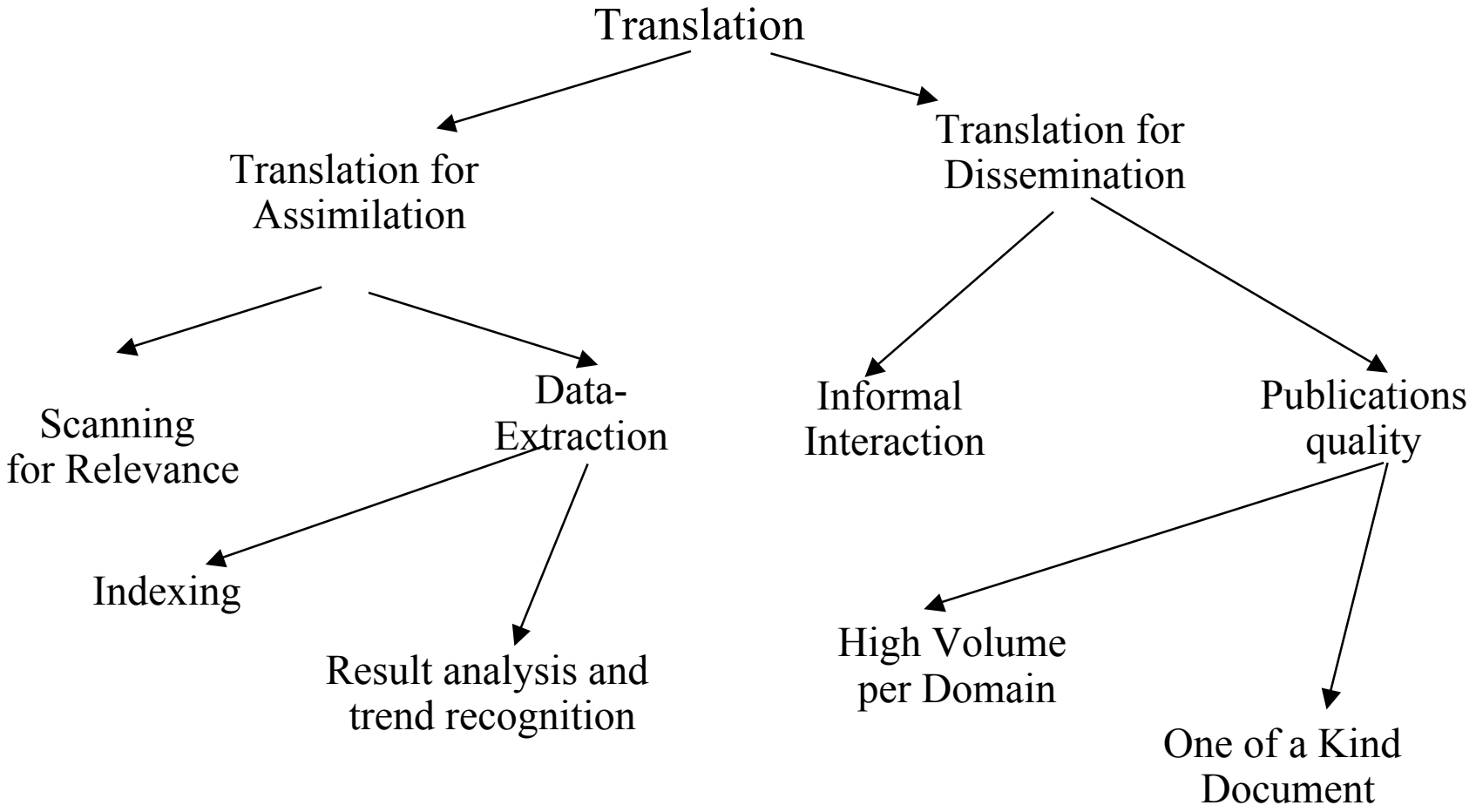
Which quality do we expect from MT systems?

Features, which we expect (at least) from MT systems:

- Semantic adequacy
- Stylistic and pragmatic adequacy
- Cultural adequacy
- Consistency inside a text and between texts
- Reduced costs compared to human translations
- High speed

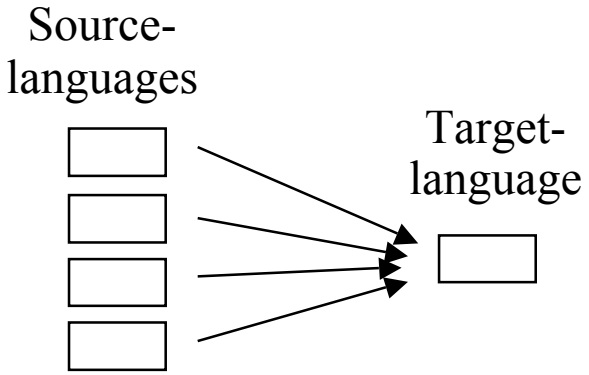
Functional Typology of MT-Systems

©Carbonell



Translation for Assimilation/Dissemination - Characteristics of the systems -

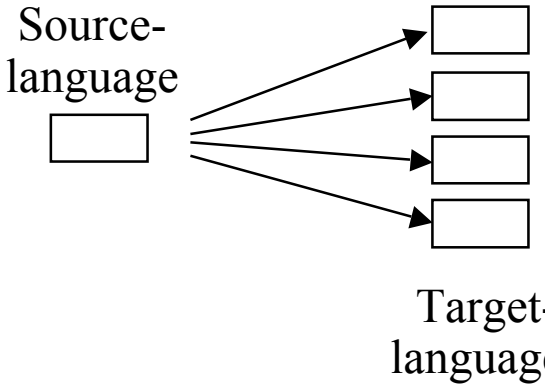
Assimilation



Process

- Each language
- Each style level
- almost each theme
- All-purpose translation
- Few Semantics
- Requires Post-editing

Dissemination



- One source language
- Style is defined
- A theme or a domain
- Special translations
- Full semantic analysis
- No Post-editing

History

Year	U.S.A	Europe	Japan
1950ies	Start of big MT-systems		Early MT-Research
1960ies	ALPAC End of MT	Start of MT	
1970ies 1980ies	(SYSTRAN, METAL) NLP Basic Reserach	GETA EUROTRA	
1990ies	Newr Start in MT-Research (SYSTRAN)	EUROTRA (METAL SYSTRAN)	MT-System MT Boom in Industry MT-Products
	Official MT-research (SYSTRAN) Multilingual Systems	End of EUROTRA NLP Basic research, VERBMOBIL	Basic research CICC, EDR, …

Machine Interpretation

- New research and technology domain with applications in
  - Consecutive interpreting
  - Simultaneous interpretation
  - Dialogue interpretation
- Interesting because of the connection between
  - Signal level ⇔ Phonetics and
  - Text level ⇔ Linguistics
- High Relevance for cognitive linguistics
  - Interpreting strategy
  - Understanding
  - Time behaviour
  - Mapping of speaker- and language features

Lexical differences between languages

- One word in a source language can be replaced (translated) through more words or multi-word expressions in the target language
- One word can be unambiguous in the target language, but not from the perspective of the source language
- Ambiguity can be found: in one language and across languages
- Lexical differences across languages have their source in
  - Difference between notions
  - Grammatical differences
  - Stylistic differences

Translation and understanding

Example: Romanian ⇒ German

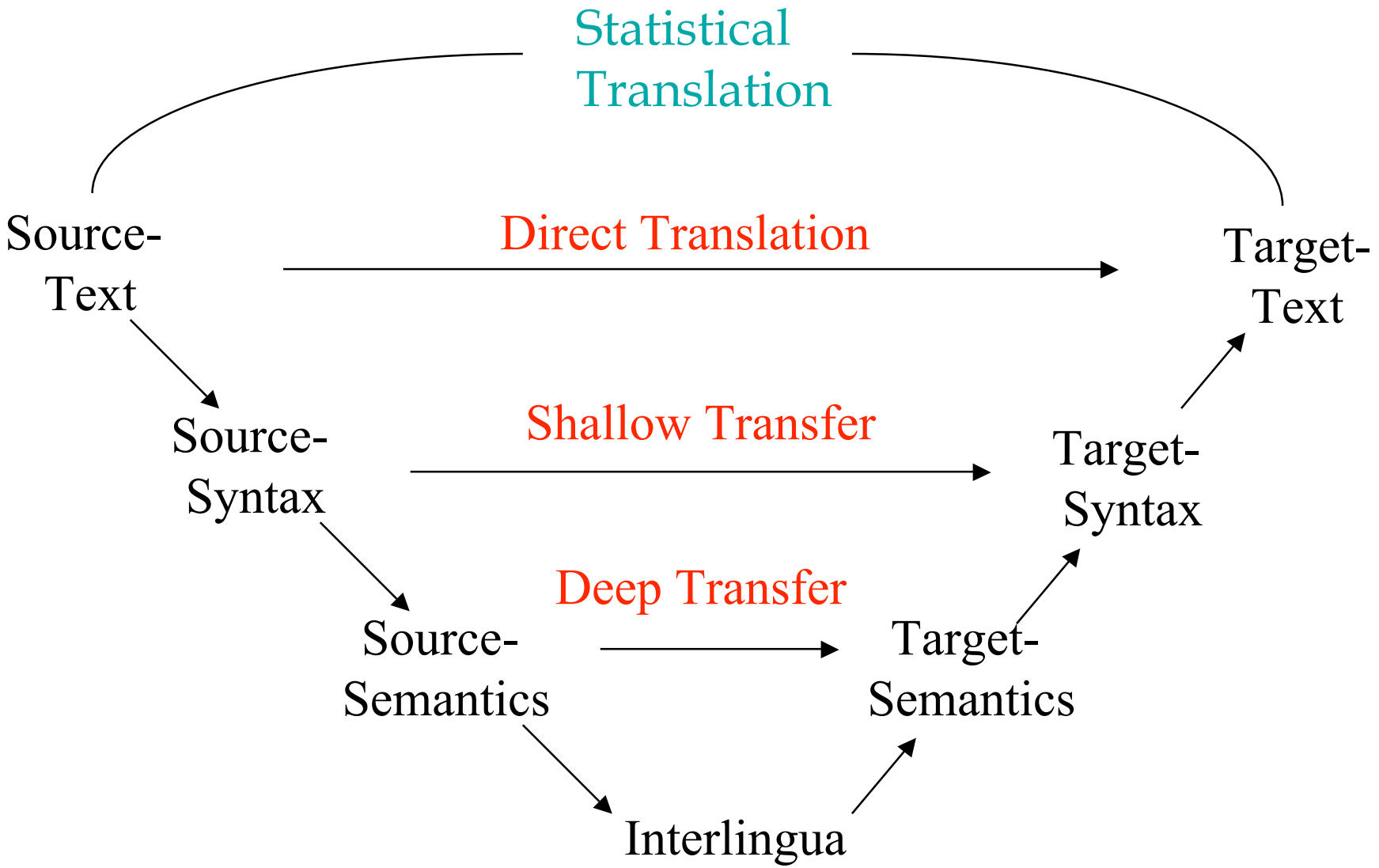
Săptămâna viitoare plecăm la Chișinău
Problem:

In German (for e.g.) "a pleca" can be translated with :

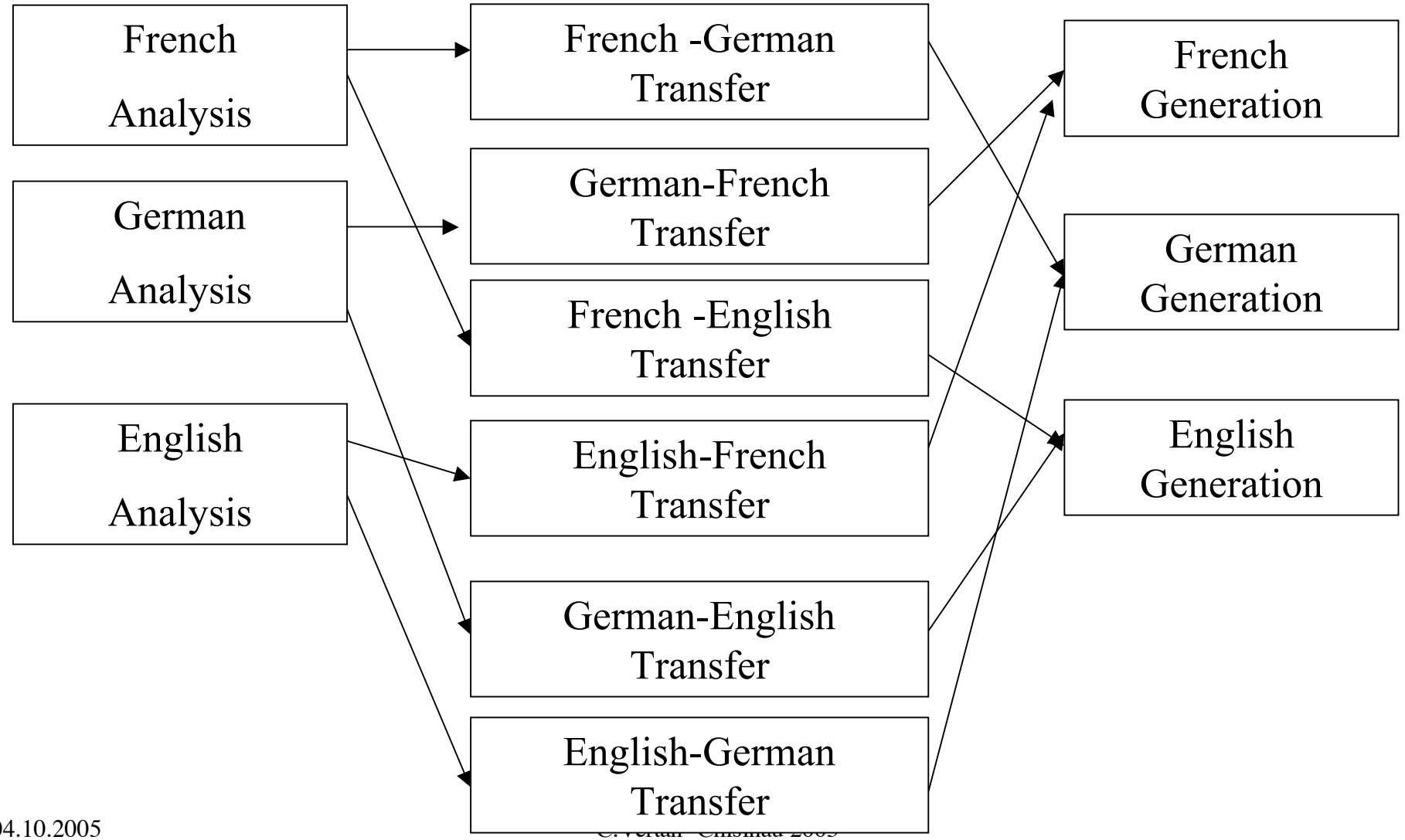
1. *gehen* to walk, go by foot
2. *fahren* to go by train, car, bicycle
3. *fliegen* to fly, go by plane

How can an MT system choose the right alternative ?
It is hopeless without (at least lexical) semantics.

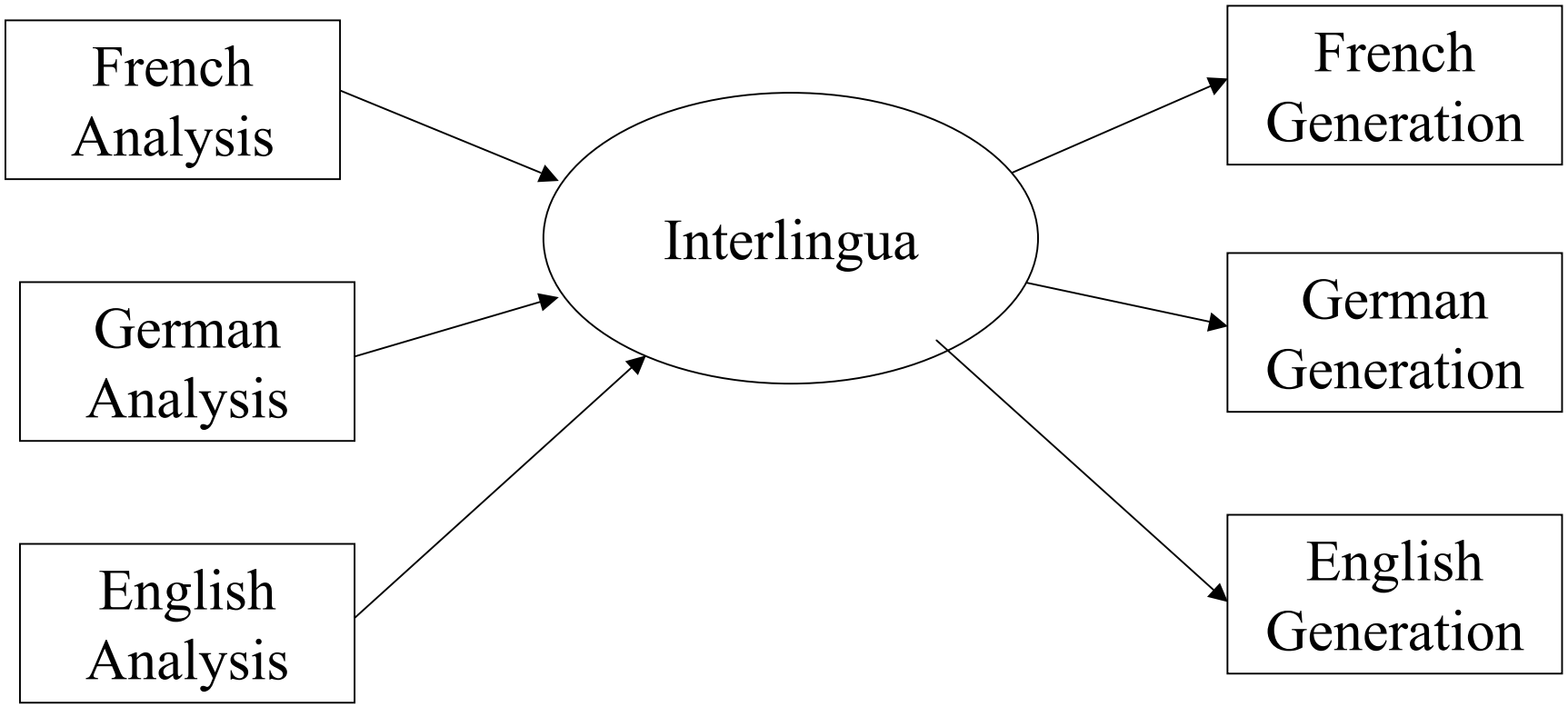
The MT-Triangle



Transfer-System with 3 Languages



Interlingua-System with 3 Languages



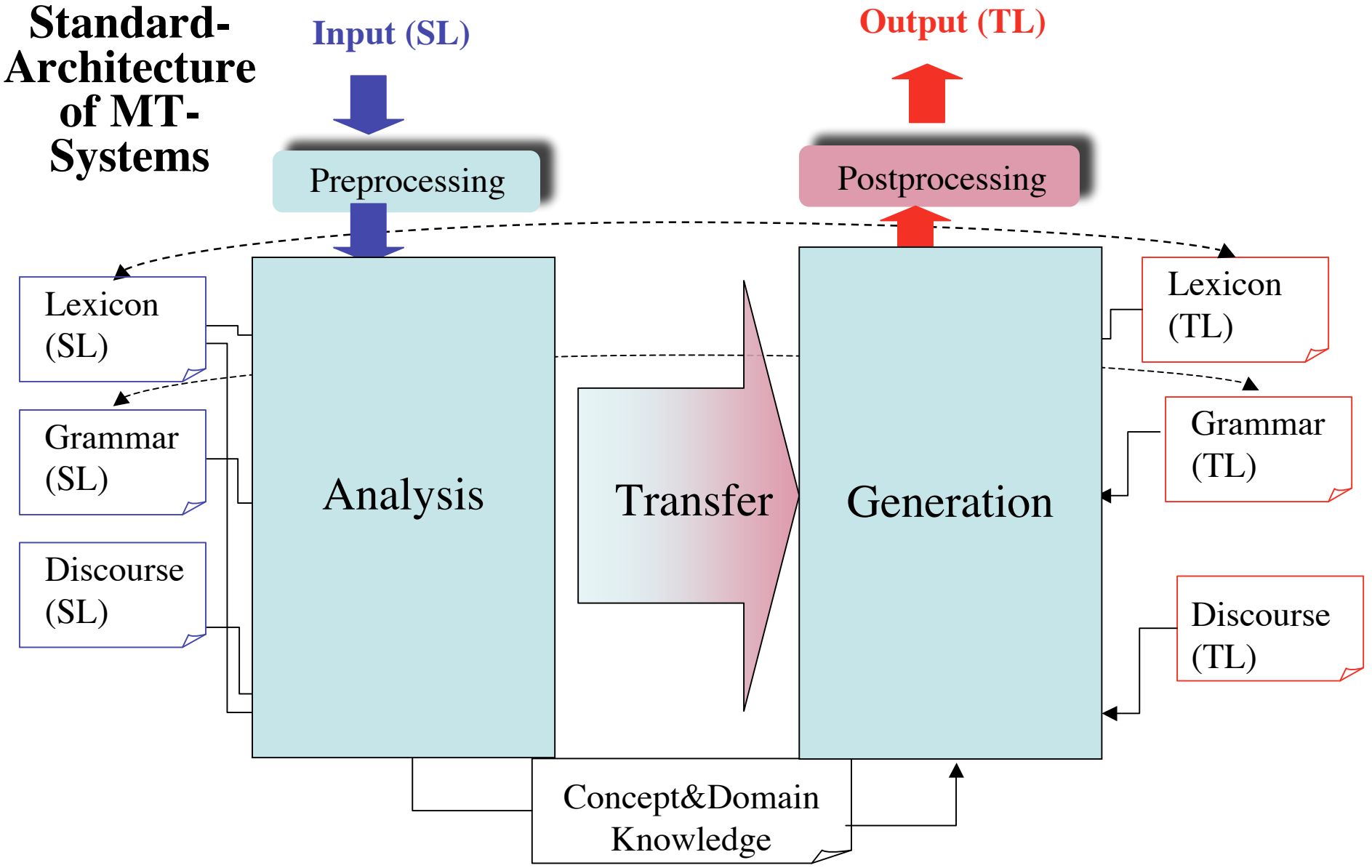
Interlingua- vs. Transfer-Systems

- Each module is independent from all other analysis and generation modules
- Target languages have no influence on the analysis process.
- For a new language only 2 new modules have to be added
- „back-translation“ possible (useful for system evaluation)
- Complicated representation even for languages belonging to the same family

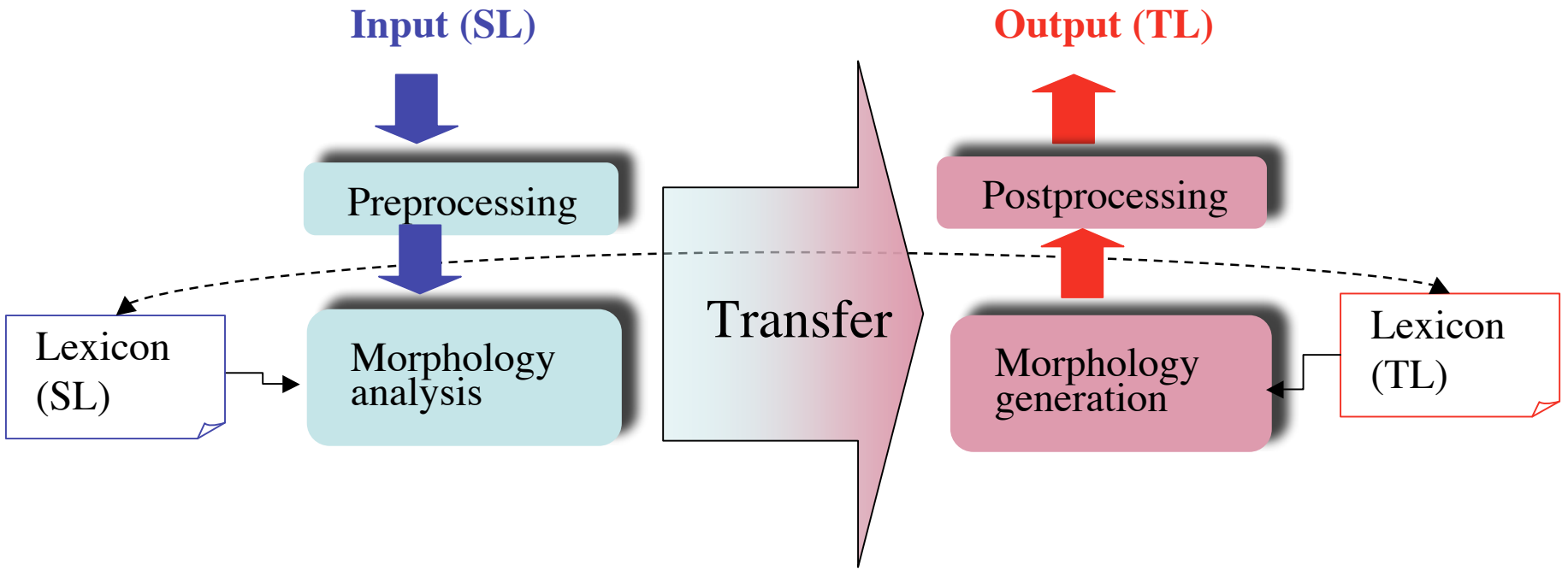
- Language-dependent
- For each new language a high number of new modules must be implemented

  (for `n` languages: `n×(n−1)` modules)
- Straight-forward representation
- Local definition of similarities among languages.

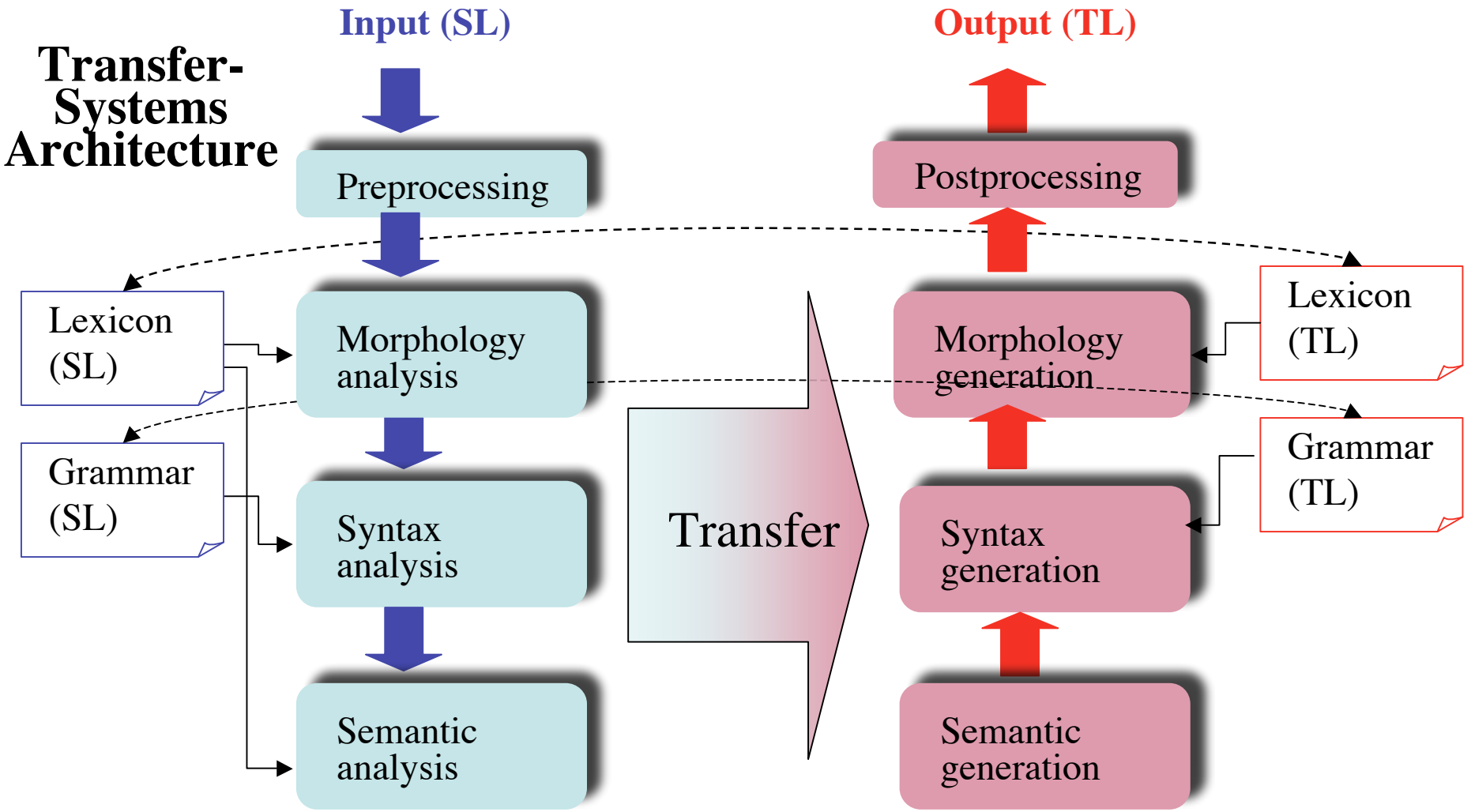
Standard-Architecture of MT-Systems



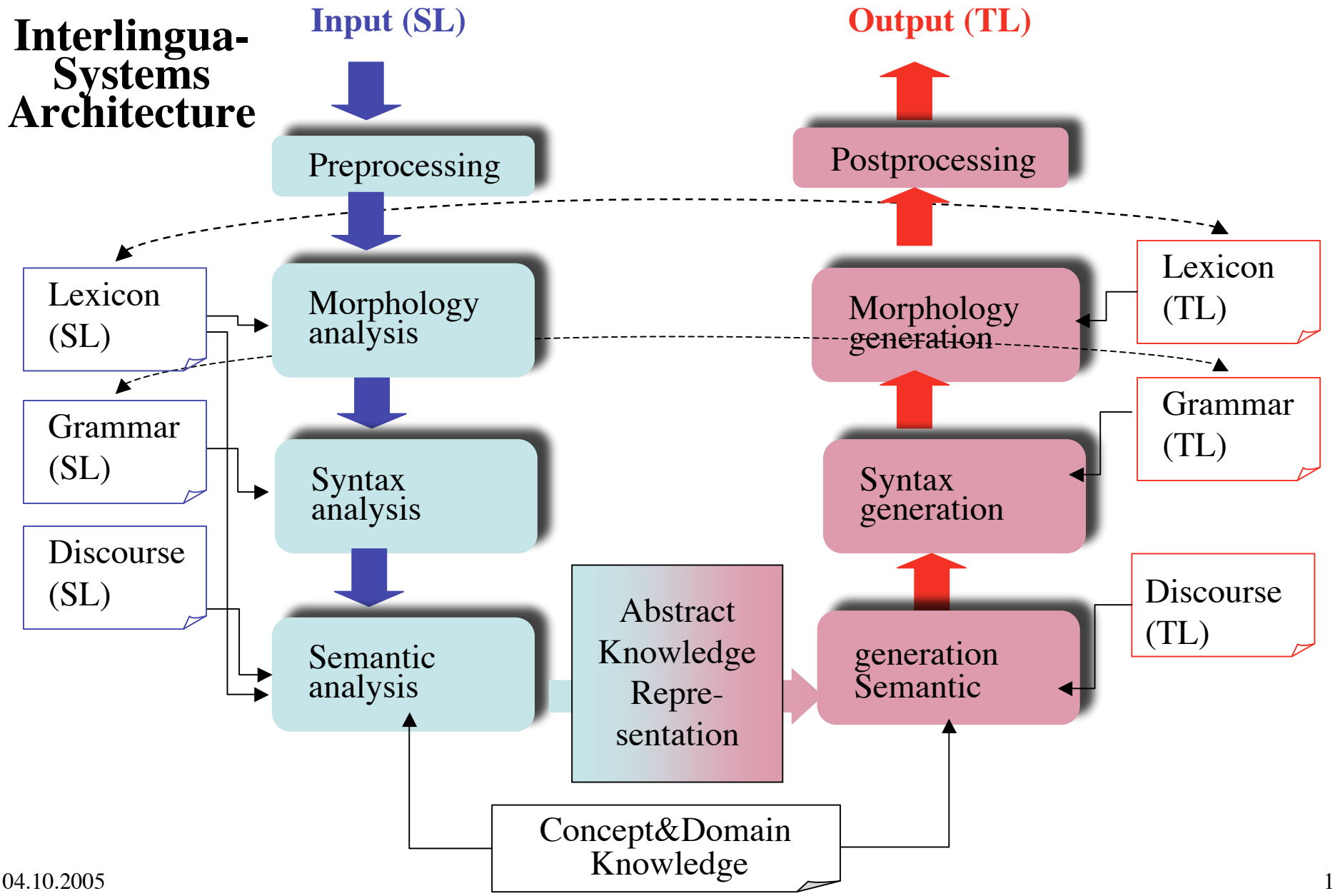
Direct System Architecture



Transfer-Systems Architecture



Interlingua-Systems Architecture



MT-specific Pre-editing

- Checking source texts for foreseeable problems for the system and trying to eradicate them
- It can include:
  - Identification of names (proper nouns)
  - Marking of grammatical categories of homographs
  - Indication of embedded clauses
  - Bracketing of coordinate structures
  - Flagging or substitution of unknown words
  - Extreme form: Reformulation of the text using a “controlled language” and a corresponding editor

Pre-editing - Controlled Language

- Adaptation of source texts to the vocabulary such constructions which the system can translate
- The writers of texts for translation are restricted to
  - particular types of constructions
  - the use of terminology,
  - predefined meanings of every-day words
- E.g the sentence: *Loosen main motor and drive shaft and slide back until touching back plate* must be rewritten into:

Loosen the main motor. Loosen the drive shaft. Slide both parts until they touch the back plate.

Post-Editing -1-

- Correction of the output from the MT-System to an agreed standard:
  - Minimal for assimilation purposes
  - Thoroughly for dissemination purposes
- E.g. Spanish ⇒ English output of an MT system:

En este studio se buscará contestar dos preguntas fundamentales

In this study it will be sought to answer two fundamental questions

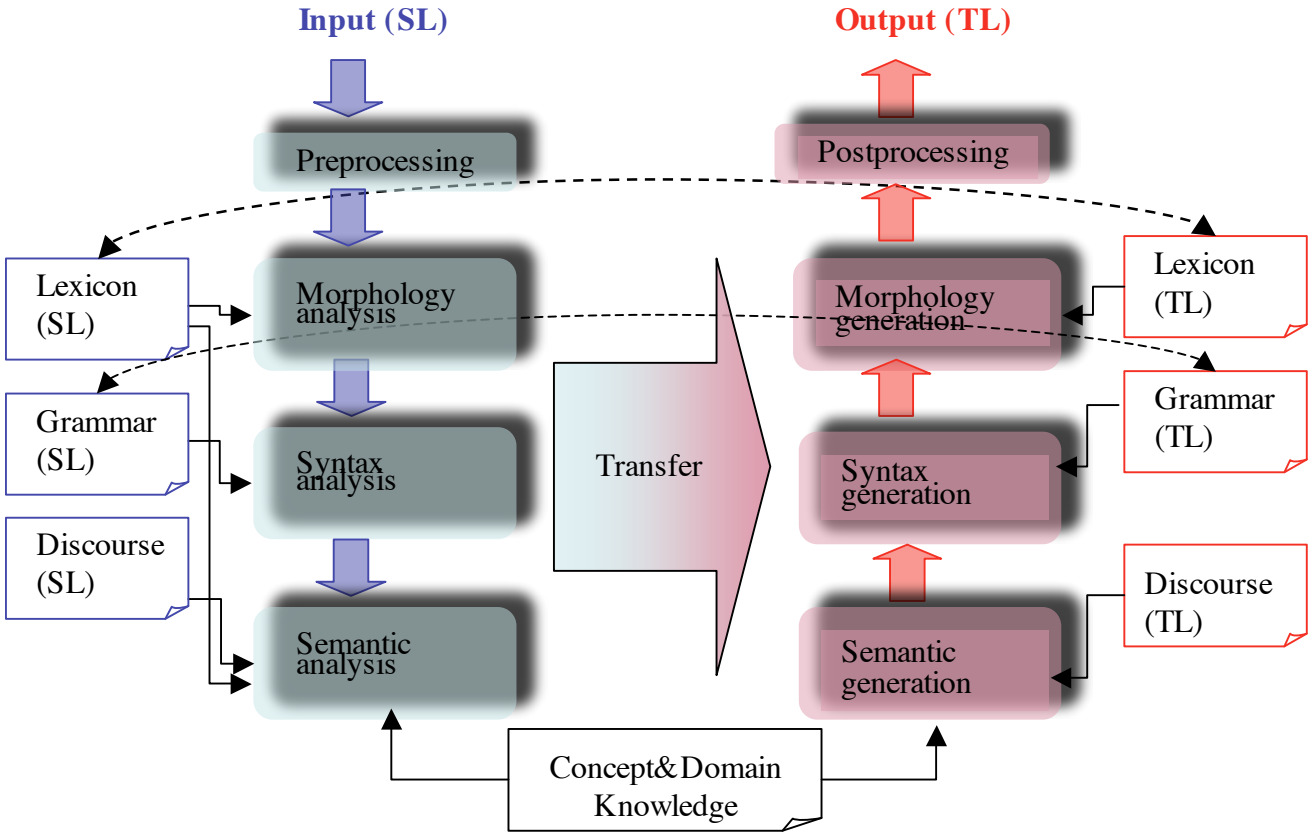
- The best post-edition may be:

This study will seek to answer two fundamental questions

Post-Editing -2-

- Interactive post-editing:
  - The system alerts the editor of sentences or phrases which may be incorrectly translated (e.g. which contain an unresolved ambiguity, or a construction which could not be analysed
  - It provides the option of correcting similar errors automatically throughout the text ,once the editor has replaced a mistranslation
- Linguistically intelligent word processors:
  - Can spot some types of structural ambiguities
  - Can generate alternative structures
  - Change automatically gender agreement in a whole phrase
  - Insert automatically appropriate prepositions (e.g if *discuss* is changed to *talk* then *about* is inserted before the direct object)

Architecture challenges for spoken language

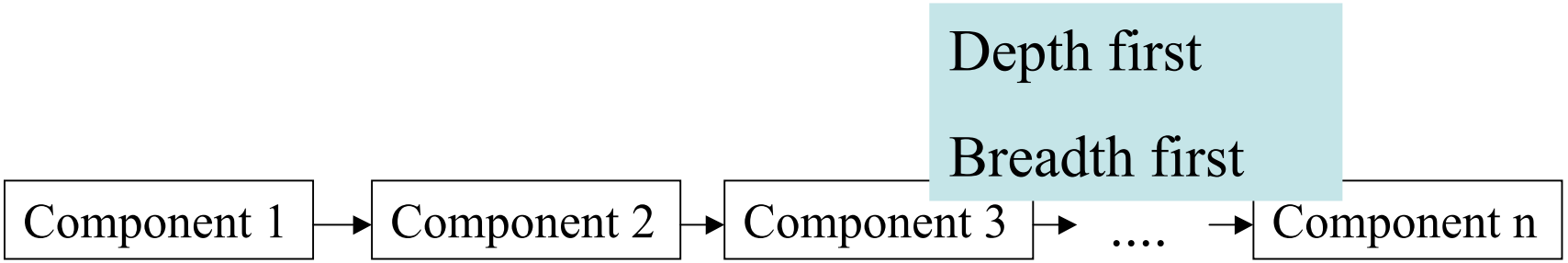


- Real - time
- incremental
- robust with un-grammatical input
- no input repetition
- usually utterances are part of a dialogue

How to connect all these modules ?

What kind of partial results should they exchange ?

Communication mechanisms: Sequential



Depth first

Breadth first

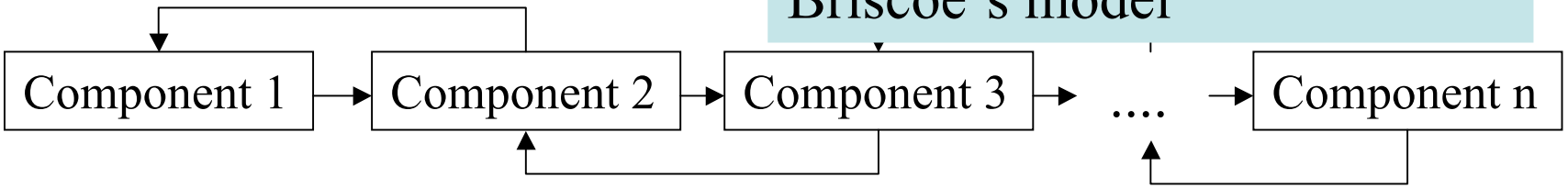
No later revision of results possible

No long distance influence among components

Communication mechanisms: Cascade

Elementary interactive model

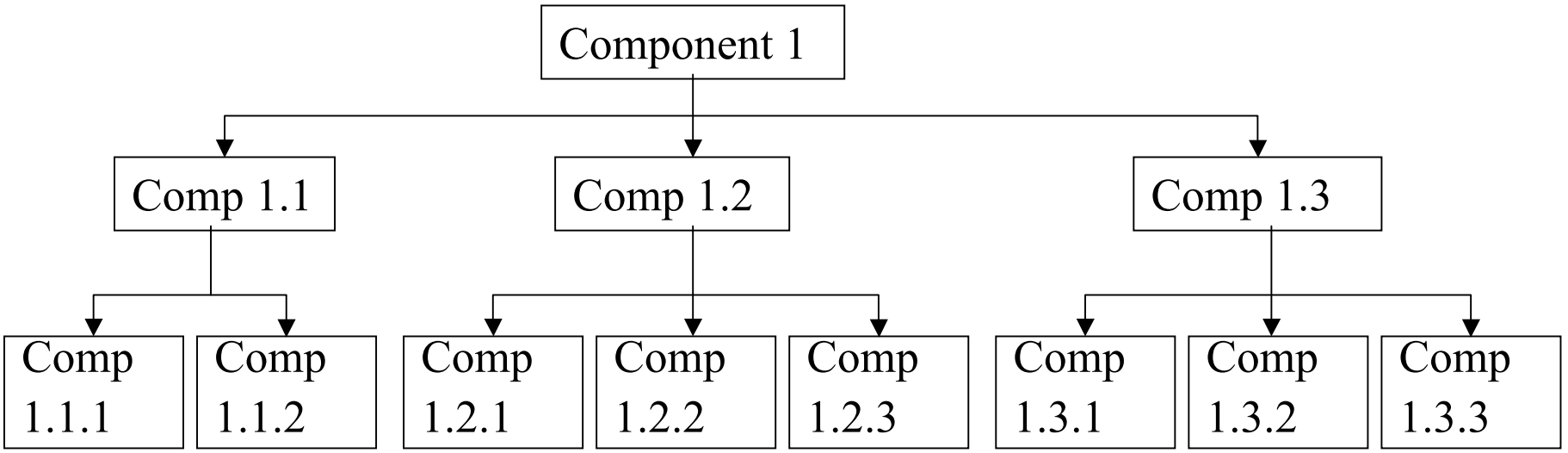
Briscoe's model



The links between modules are static, no other connections are permitted

Classical Communication mechanisms:

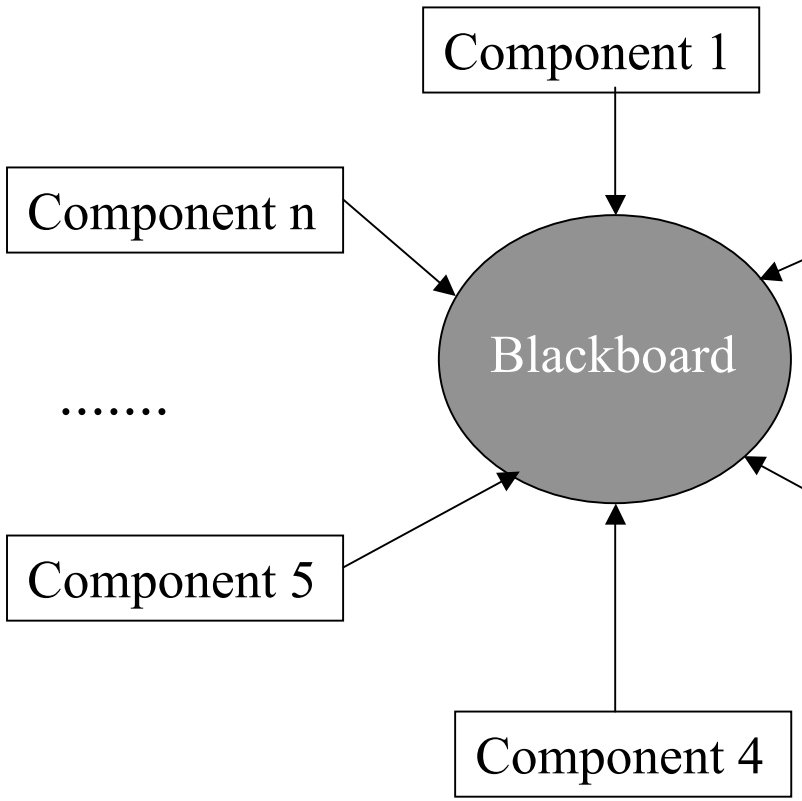
Tree structure (sequential multi-layered)



Communication mechanisms: Blackboard

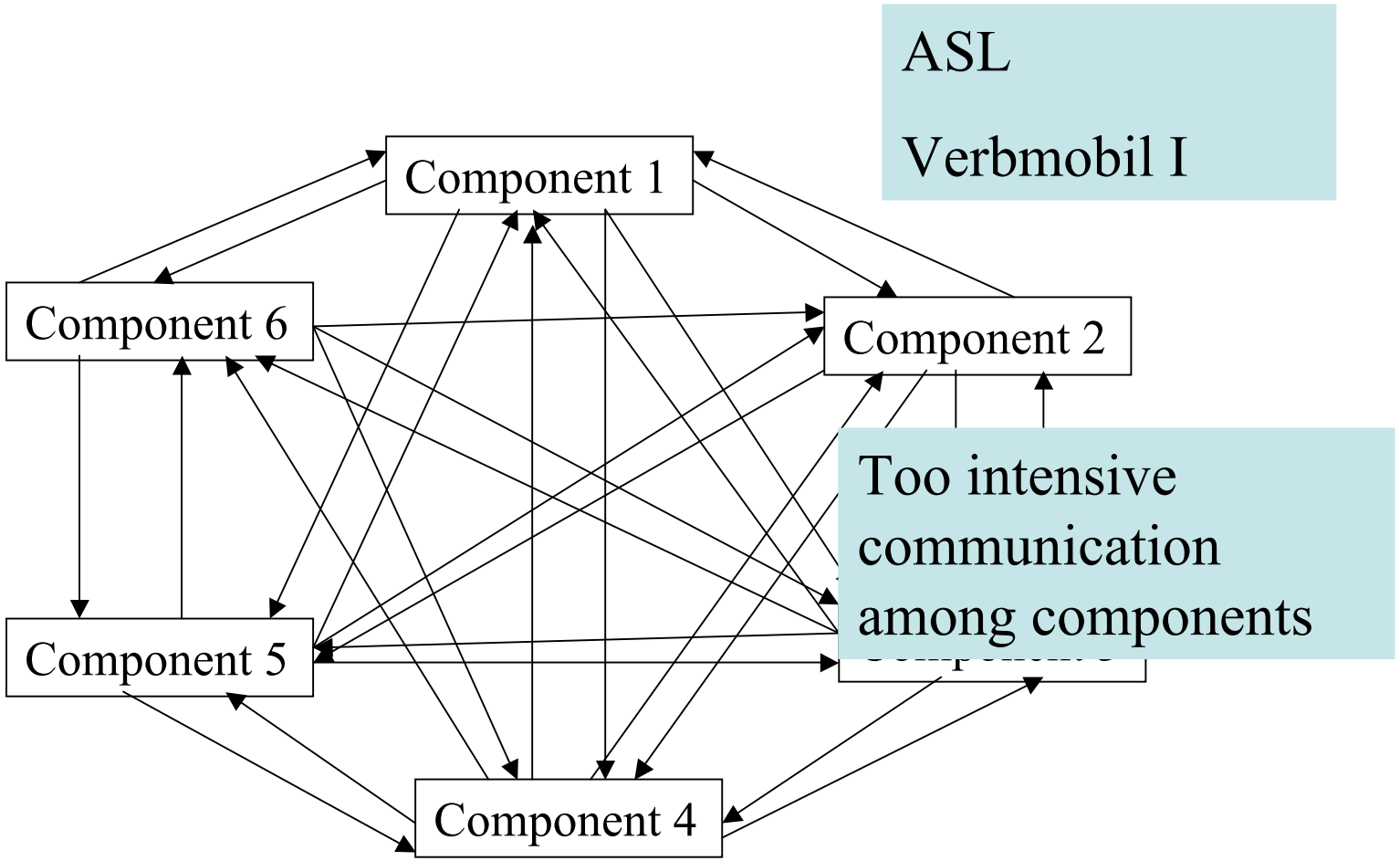
HEARSAY

Verbmobil II



Bottlenecks due to central scheduling

Communication mechanisms: Multi-agent



Different Approaches to MT

- Rule-based MT
- Knowledge-based MT
- Statistical-based -MT
- Example-based

Other approaches to computer assisted translation

- Machine Aided Translation
- Translation Memories