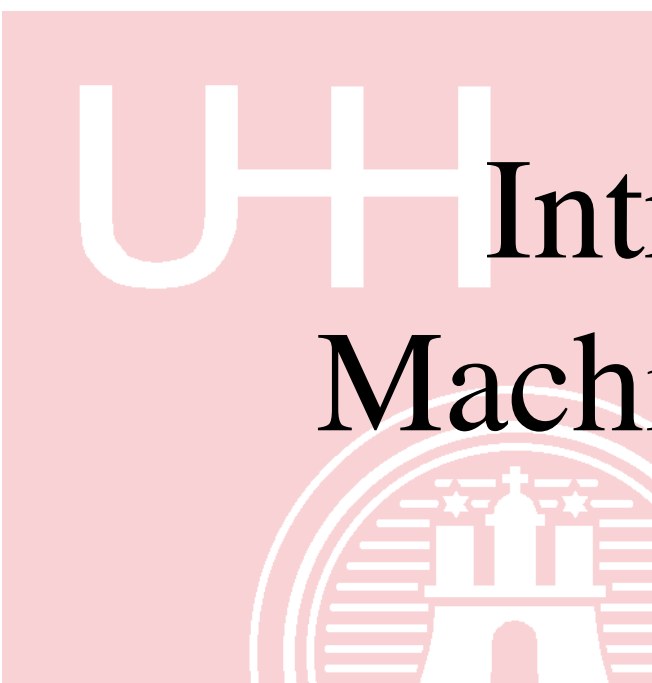


The logo of the University of Hamburg, featuring a stylized 'UH' in white on a red background, with a circular emblem below it containing a building and stars.

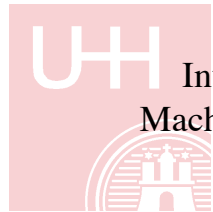
Introduction to Machine Translation

Cristina Vertan

University of Hamburg • Informatics Department
Natural Language Systems Group

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

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



Introduction to Machine Translation

Cristina Vertan

University of Hamburg • Informatics Department
Natural Language Systems Group

© C. Vertan ESSLI 2004

1

- Part I: Basic Methods of Natural Language Processing for MT
- Part II: Basic Concepts and Technologies of Machine Translation
- Part III: Rule Based Machine Translation
- Part IV: Knowledge Based Machine Translation
- Part IV: Corpus Based Machine Translation

© C. Vertan ESSLI 2004

2



Basic Methods of Natural Language Processing for MT

Cristina Vertan

University of Hamburg • Informatics Department
Natural Language Systems Group

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

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

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

© C. Vertan ESSLI 2004

3

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Why starting with NLP

- MT has much in common with all other language technologies:
 - The basic features of natural languages
 - Missing situational data
 - Limited modal channels
 - Limited technology
 - New words/names
 - Non-deterministic processing

© C. Vertan ESSLI 2004

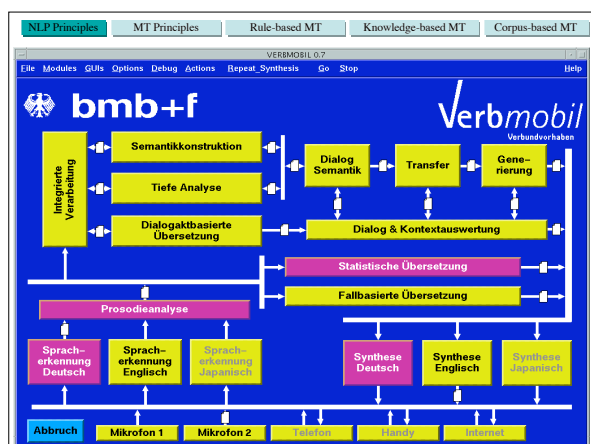
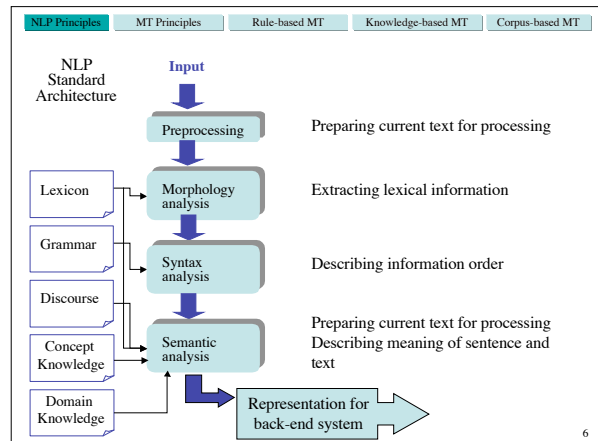
4

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Typical Features of Natural Language

- Unclear focus of analysis, esp. with spoken input (the whole text only?)
- Ambiguity on all levels
- Self-reference, meta language capacity (I meant "lion")
- Valencies, i.e. syntactic/semantic co-occurrences of categories
- Multi-word lexemes and idioms with non-compositional meaning
- Hierarchical syntax in non-linear order
- Long distance dependencies
- Discontinuous components (ese ...ah!)
- Ellipses (Imagine these phenomena in programming languages ...)
- Paraphrases
- Coherence
- Understanding by word knowledge

© C. Votaw ESSL1 2004 5



NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Some Central Decisions of Analysis

- Which type of language is expected?
 - Spoken input may contain errors
 - Spoken language style (even in written transcripts) may have syntactic and semantic "errors"
 - Written language has no prosody, but is supposed to be correct
- Which type of output?
 - Table, representation expressions, slot filler, classification
- Which domain?
 - Technical, social, leisure,
- Which pragmatics?
 - Question answering, action control, information,

© C. Votaw ESSL1 2004 8

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Result: GUS

Semantic Slot Fillers

"I want to go to San Diego on May 28"

↓

(Client Declare

(Case for want / e (Tense Present)

Agent = Dialog.Client.Person

Event = (Case for go (Tense Present)

Agent = Dialog.Client.Person

To-Place = (Case for City

Name = San Diego)

Date = (Case for Date

Month = May

Day = 28))))

Winograd

© C. Vortan ESSLI 2004

9

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Result: DB-Interface

Expression

"List the names of all suppliers, who deliver at least the parts that are delivered by supplier S2"

↓

SELECT

UNIQUE S#

FROM

SP SP X

WHERE

NOT EXISTS

(SELECT *

FROM SP SP Y

WHERE S# = 'S2'

AND

NOT EXISTS

(SELECT *

FROM SP

WHERE S# = SP X. S#

AND P# = SP Y. P#))

© C. Vortan ESSLI 2004

Date

10

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Result: Table of chemical reactions in SIE

REF para. 1	SCALE small	PHASE solid	YIELD 77%	TEMP -78 to 20
TIME	ENERGY cooling	APPARATUS	FEATURES IR. NMR. MS	
REG. NO	FUNCTION	AMT.	AUTHOR ID	
78624-62-1	product	2.70 g	7a	
78624-61-0	reactant	1.24 g	6a	
13274-48-6	reactant	80 ml	N-methyltriazolinedione	
	solvent	40 ml	pentane	
			ethyl acetate	

© C. Vortan ESSLI 2004

11

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Result: Semantic Representation

[request (referent(_5747))

presuppose (exists (_4340))]

some (_4340)

[unique (_4407)

single (_4407)

instance (_4407, person)

propval (person_4407.sex,male)

[some (_4725)

[unique (_5033)

single (_5033)

instance (_5033.project)

propval (project_5033.name,str (LOKI))]

"who is the man that leads the LOKI project?"

↓

instance (_4725,leading)

propval

(leading_4725,theta_5033)

propval

(leading_4725,alpha_4407)

topic (_4407)]]

instance (_4340,identity)

propval (identity_4340.alpha_4407)

propval (identity_4340.theta_5747)

topic (_4407)

[some (_5747)

single (_5747)

instance (_5747.person)]

© C. Vortan ESSLI 2004

12

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

First Step: Preprocessing

- Separate text from non-text (images, code, analyze tables ...)
- Lemmatization (Splitting)
- Normalize writing (e.g. Ablaut)
- Join separable suffixes (esp. in German)
 - "Er fing die Maus ein" → *erfangen*
 - (*He caught the mouse in)
- Separation of compounds
- Block multiword terms and idioms,
- (Attach PoS

Lemmatizer

Tokenizer

PoS-Tagger)

© C. Vortan ESSLI 2004 13

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Second Step: Consulting the Lexicon

- Full form lexicon
 - Every form of a word is an entry in the Lexicon
 - Result: no morphological processes after tokenizing
 - For real life applications sometimes too large
 - Difficult for languages with strong composition
- Stem lexicon
 - Only stems are entries
 - Additional information about inflexion class
 - a morphological generator is necessary
 - Result: small resources

All further processes, except for names, rely on lexicon information, at least on part of speech tags

© C. Vortan ESSLI 2004 14

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Resources: Lexicon

- **Dictionary**
 - Pronunciation
 - Definitions
 - etymological information
 - stylistic information
 - PoS (part of speech)
 - Few sub-classification features (usually gender, plr.)
 - Translation (in bi- or multilingual dictionaries)
- **Lexicon**
 - PoS (noun, Verb, etc.)
 - Sub-classification features (verb transitive/intransitive, Genus etc.)
 - Inflexion classes
 - semantic information (e.g. if a verb requires an alive Subject)
 - a link to translation equivalents in other lexicons, or a mark for lexical gap (in bi- or multilingual dictionaries)

© C. Vortan ESSLI 2004 15

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Resources: Lexicon

Representation and Encoding - 1 -

- There is a huge number of lexicon formats according to:
 - the encoded linguistic information (which features, in which order)
 - the encoding schema (distributed lexicon, delimiters between linguistic categories or entries, pointer to entries in other lexicons)
- The lexicon design is an extremely time consuming process, therefore „re-usability“ has high priority on the agenda of lexicon developers.,
- Several standard models have been proposed (PAROLE / SIMPLE, MILE) as well as standard encoding schemas based on XML (SALT, OLIF)

© C. Vortan ESSLI 2004 16

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Resources: Lexicon

Representation and Encoding - 2 -

- Most existing standard models are very complicated because they intend to cover a large spectrum of linguistic features but
 - they still did not succeed to model all linguistic phenomena of the European languages,
 - Existing lexicons, which do not follow these standards, cannot be re-used.
 - MANAGELEX (Univ. Hamburg) is a tool for reformatting, editing, developing and merging of lexicons (under development)

© C. Vortan ESSLI 2004

17

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Pattern matching - 1 -

- Pattern = a syntactic frame for lexical-semantic equivalence classes.
- Patterns describe frequent expressions of a language

Wo is president of Germany ?
 mayor Granada
 king Denmark

© C. Vortan ESSLI 2004

18

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Pattern matching - 2 -

- Patterns specify
 - A fixed word order
 - Places for variable words
 - A clause can be processed, if there is (at least) one matching pattern
- Patterns can be filled by categorical, syntactic or semantic constraints
 - "_any word_" / "_any noun_" / "_any noun+sing_" / "_any noun with attr+living_" / "any noun class MUSHROOMS_"
- For one sentence more than one pattern may be applied

Problems: syntactically or semantically deviant sentences can be accepted, if the equivalence classes are not defined tightly:

- z. B: **Who is** the pencils **of** football league ?

Maintenance problems with large number of (partly overlapping) patterns

➡ Patterns do not deliver a symbolic description

© C. Vortan ESSLI 2004

19

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Basic Syntactic Decisions

• System type	Parser&Grammar vs	separate grammar modules
• Result of parser	Full vs	selective
• Type of grammar	Dependency vs	constituency
• Formalism	any sort	
• Architecture	parallel vs	sequential
• Start point	top-down vs	bottom-up
• Rule application	deterministic vs	non-deterministic
• Rule choice	first guess vs	informed choice
• Strategy	breadth first vs	depth first
• Scope	word by word vs	phrase by phrase
• Ambiguity handling	until success vs	exhaustive

© C. Vortan ESSLI 2004

20

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Syntactic analysis

- The output of the morphological analysis is parsed according to the chosen grammar
- i.e: parsing of a sequence of PoS symbols (retrieved by the morphological analysis)
 - the correct order of the PoS is proved
 - Iteratively a structural description is written into a data structure
 - <Structure= Art + N +>
- Very often a part of the input is abandoned for the moment because substructures have to be analyzed first.
- e.g *Los libros, que los hay comprando ayer, son muy interesante*
 <Art + N+ <Colon + sub-clause...> + V +Mod+Adj
 <Rel +Pron +V+Adv>

© C. Vortan ESSLI 2004

21

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Resources: Grammars

- Grammars define the conditions of well-formed expressions in a language (syntax)
- Describe three basic relationships in sentences:
 - sequence of words* (in English adjectives normally precede the nouns that they modify, whereas in Spanish they normally follow it)
 - categories*: e.g. a noun phrase may consist of a determiner and a noun or a determiner, an adjective and a noun.
 - dependency* i.e. relations between categories: prepositions determine the case of the nouns which depend on them: e.g. „mit“ (germ.) „con“ (sp.) always require dative

© C. Vortan ESSLI 2004

22

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Resources:Grammars

Grammar types

- Two basic types of grammatical representations are in common use, dependency grammars or constituency grammars.
- Sequence is optionally indicated in both types of representations
- Dependency relations are represented
 - in a dependency grammar by a word tree starting with the verb
 - In a constituency grammar by a tree of constituents
- Categories are explicitly represented in a phrase structure tree, in s dependency grammar categories are subtrees

© C. Vortan ESSLI 2004

23

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Resources: Grammars

- Example -

S[] NP VP

NP[] N

NP[] NP AdjP

NP[] Det N

AdjP[] Mod Adj

VP[] VNP

N[] Granada, ciudad

V[] es

Det [] una

Mod[] muy

Adj[] bonita

Constituent structure

Dependency structure

Granada es una ciudad muy bonita

© C. Vortan ESSLI 2004

24

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Resources: Grammars

Feature -based representations

- To shift grammatical features to higher nodes or inherit features from them
- Linguistic features are represented as attribute-value-pairs
- Additionally, rules for combining features must be specified (e.g. for correspondence)
- Features can be inserted both in constituent and dependency structures
- E.g.
NP [Gender,Nr] Det [Gender, Nr] N[Gender,Nr]

© C. Vortan ESSLI 2004 25

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Basic Semantic Decisions

- Semantic inside the parser? □ semantic parser / case grammar
- Separation of construction, resolution and evaluation? □ Multi-phase processing vs extraction
- Frame oriented processing or compositional treatment
- Domain knowledge in semantics? □ Reference semantics
- Conceptual knowledge separated from facts? □ interaction
- User specific interpretation? □ Partner model
- Time-dependent? □ time logic

© C. Vortan ESSLI 2004 26

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Semantic Strategies

- Elementary: Key word spotting
- Basic: Syntax looks only for semantic slot fillers
- Technical solution: The parser delivers already a semantic structure by looking for semantic roles and dependencies only
- Standard: Lexical semantic entries are amalgamated with parsing result **Semantic case grammar**
- Advanced: A full logical representation of the proposition and presupposition is built up

© C. Vortan ESSLI 2004 27

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Semantic Roles /Deep Case Semantics

Based on lexical semantics and syntax, sentence semantics delivers the semantic **potential** of an utterance

Often used: Semantic Roles (Deep Cases):

- Actor
- Instrument
- Object, etc.

specify "persons and items" of a sentence,

E.g., in an action:
An <Actor> moves an <Object> from a <Location1> to a <Location2> along a <Path> for a <Beneficiary>.

Grammarians propose up to 25 roles. Specific domains may have a very limited number of roles (e.g. weather reports)

© C. Vortan ESSLI 2004 28

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Semantic Resolution

Define the current meaning by

- Using contextual knowledge:
 - Determination of current values
 - Disambiguation
- Using domain and every-day knowledge:
 - Identification of referred objects
 - domain data and state of affairs
 - Determination of relevant utterances
 - User specific inferences

Discourse Semantics

Reference Semantics

© C. Vortan ESSLI 2004 29

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Resources:

World Knowledge and Domain Knowledge

- Concepts and relations between concepts are represented in an ontology (semantic feature hierarchy)
- The lexemes (lexicon entries) in different languages can be mapped onto this ontology (details in knowledge-based MT)
- Sensors or time dependent expressions describe the state of affair (speaker t_1 = Antonio = "I")
- Usually this information is language independent.

© C. Vortan ESSLI 2004 30

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Ambiguities on all levels

The central difference between formal and natural languages is the ambiguity. E.g.,

- Speech ambiguity "Lead a ship" vs. "leadership"
- Lexical ambiguity "peak" vs. "peek"
- Syntactic ambiguity "Drive to the bank, please!"
- Syntactic ambiguity "I saw the Grand Canon flying to New York"
- Pragmatic ambiguity "Can I print some reports?"
- Referential ambiguity "He took some papers out of the envelopes and send them to his boss"

© C. Vortan ESSLI 2004 31

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Discourse Coherence

So far, the scope of analysis was the sentence. However, many syntactic and semantic structures are super-segmental, especially in spoken language and in spoken style.

Important tasks: Anaphora resolution "He wanted all of the items"

Kataphores "John did the following:"

Ellipses "and those for 2005?"

Pronouns can often be resolved (replaced by their antecedent) by searching for an adjacent noun in the previous sentence, which has the same grammatical and semantic features (role restrictions of the verb)

Kataphores are much more difficult, but they are rare.

Ellipses can be completed by testing unification with previous sentences

© C. Vortan ESSLI 2004 32

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Resources: Discourse memory

Minimal case:

- List of before-mentioned objects with gender information (for easy pronoun resolution)

Best case

- List of before-mentioned objects with gender information and semantic features (for elaborated pronoun resolution)
- Memory of syntactic structures (for ellipses reconstruction)
- Memory of propositions (for ellipses reconstruction)
- Memory of speech acts (for pragmatically adequate reactions)

© C. Vortan ESSLI 2004 33

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Pragmatics reasoning

- Pragmatics is the linguistic field, which describes relations between language and action (planning). E.g.:
- Whenever I say "can you do X?" I assume, that
 - it is an order, not a yes/no-question
 - I want X
 - X can be done
 - I want it to be done immediately
 - I expect a rejection in case of disagreement
 - I am responsible for X, etc

All pragmatic presuppositions and implicatures must be fulfilled in the discourse. This is especially important with legal and economic translation. Counterexample:

"Can you ship these trains?" ☐ "yes, we do it immediately!"

© C. Vortan ESSLI 2004 34

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Generation

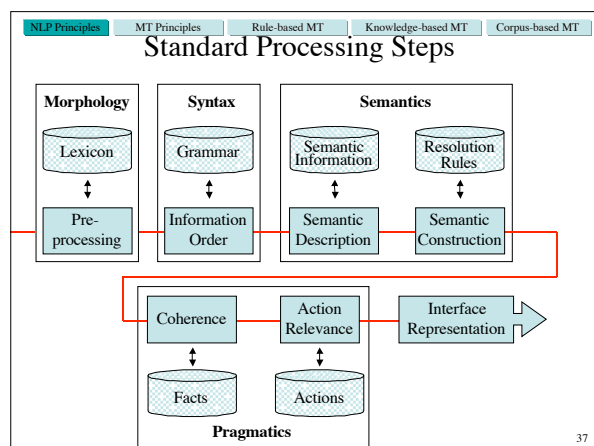
- Elementary solution in simple domains: Output patterns with the result values:
- "Who is the president of the UN ☐
- "Kofi Annan"
- Better solution: The syntax for the answer is extracted from the question Processes:
 - Consider topicalization
 - Adjust word order
 - Omit trivial parts to be not too repetitive
- Best solution: Evaluation renders propositions (predicate-argument structures) and adequate grammar rules allow for free generation.

© C. Vortan ESSLI 2004 35

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Thank you!

© C. Vortan ESSLI 2004 36





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: cri@informatik.uni-hamburg.de

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

NLP Principles **MT Principles** Rule-based MT Knowledge-based MT Corpus-based MT

Abbreviations

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

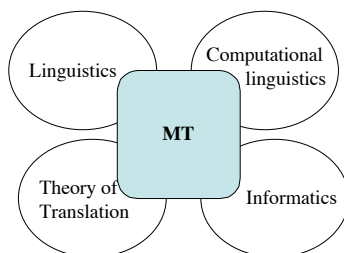
C.Vertan -ESSLI 2004

2

NLP Principles **MT Principles** Rule-based MT Knowledge-based MT Corpus-based MT

Machine Translation as a Discipline

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



C.Vertan -ESSLI 2004

3

NLP Principles **MT Principles** Rule-based MT Knowledge-based MT Corpus-based MT

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!

C.Vertan -ESSLI 2004

4

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

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

C.Vertan -ESSLI 2004

5

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Functional Typology of MT-Systems

©Carbonell

```

graph TD
    Translation --> TranslationForAssimilation[Translation for Assimilation]
    Translation --> TranslationForDissemination[Translation for Dissemination]
    TranslationForAssimilation --> ScanningForRelevance[Scanning for Relevance]
    TranslationForAssimilation --> DataExtraction[Data-Extraction]
    ScanningForRelevance --> Indexing[Indexing]
    DataExtraction --> ResultAnalysis[Result analysis and trend recognition]
    TranslationForDissemination --> InformalInteraction[Informal Interaction]
    TranslationForDissemination --> PublicationsQuality[Publications quality]
    InformalInteraction --> HighVolume[High Volume per Domain]
    PublicationsQuality --> OneOfAKind[One of a Kind Document]
  
```

C.Vertan -ESSLI 2004

6

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Translation for Assimilation/Dissemination - Characteristics of the systems -

Assimilation

Source-languages

Target-language

Process

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

Dissemination

Source-language

Target-languages

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

C.Vertan -ESSLI 2004

7

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

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	(SYSTRAN, METAL) NLP Basic Reserach	GETA EUROTRA	
1980ies			
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, ...

C.Vertan -ESSLI 2004

8

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

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

C.Verlan -ESSLI 2004 9

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

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

C.Verlan -ESSLI 2004 10

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

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

C.Verlan -ESSLI 2004 11

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

Translation and understanding

Example: English Spanish

While driving down route 72, John swerved and hit a tree

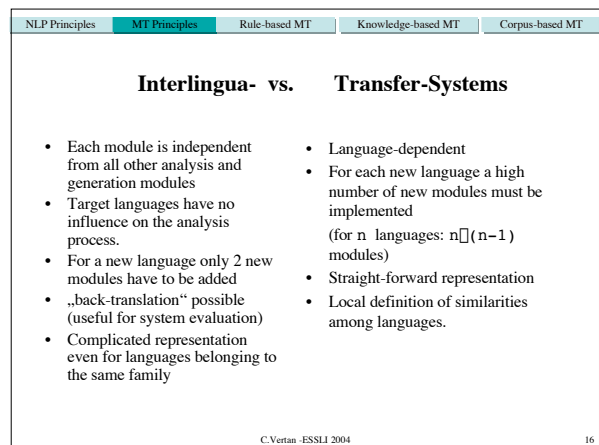
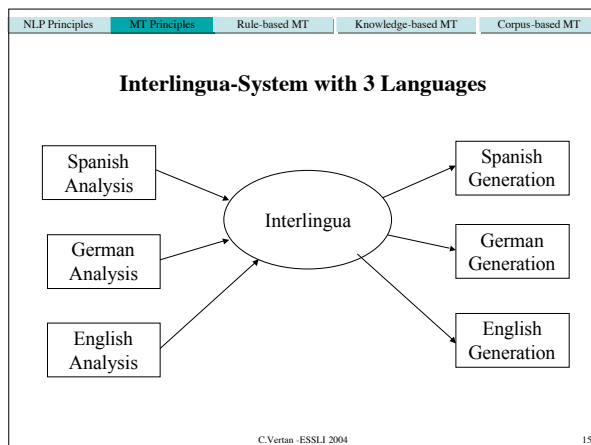
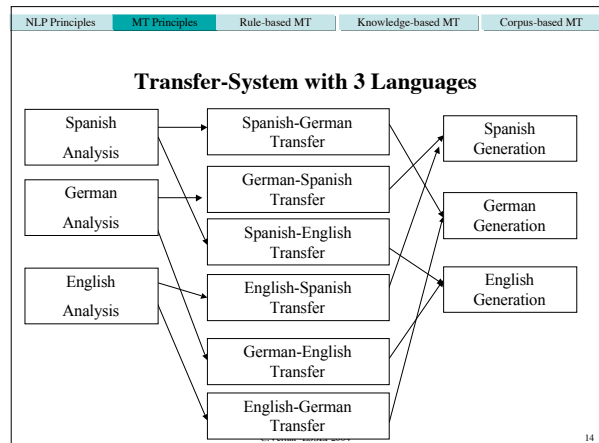
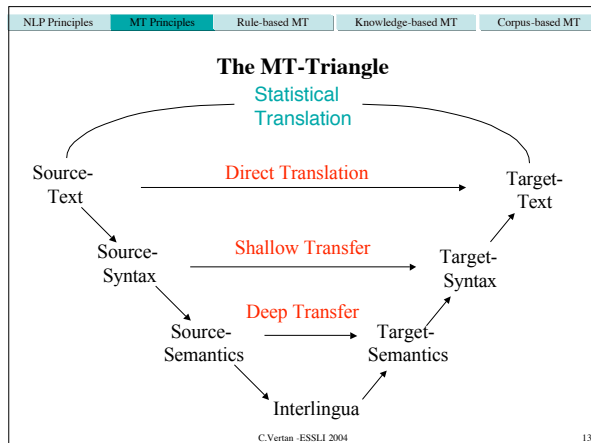
Problem:

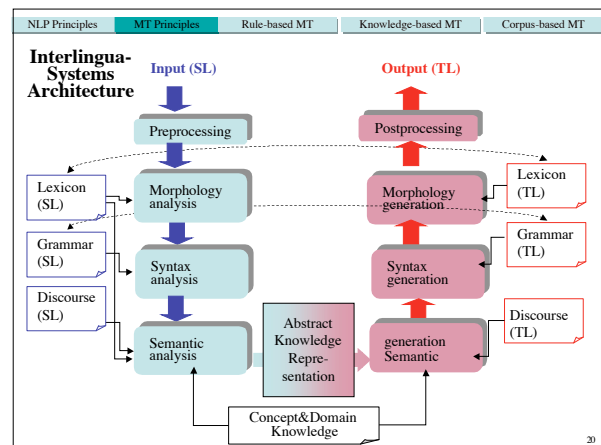
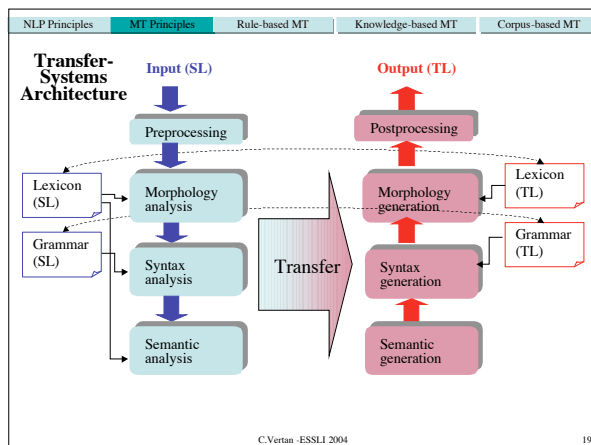
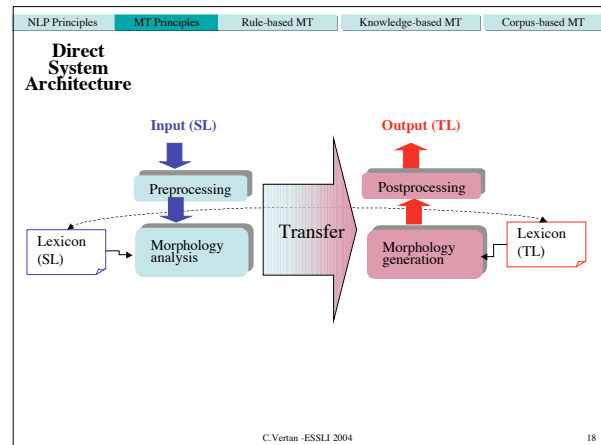
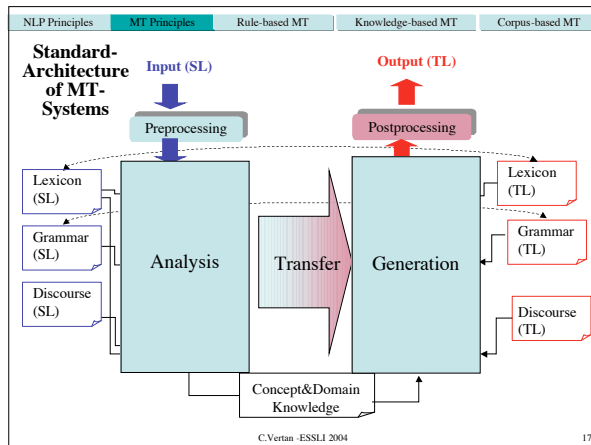
In Spanish “hit” can be translated with :

1. *pear* to move with intention
2. *chocar* accidentally due to a moving object
3. *acertar* guessing
4. *golpear* (other possibilities) etc.

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

C.Verlan -ESSLI 2004 12





NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

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

C.Vertan -ESSLI 2004 21

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

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.*

C.Vertan -ESSLI 2004 22

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

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

C.Vertan -ESSLI 2004 23

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

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)

C.Vertan -ESSLI 2004 24

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

Evaluation of MT-Systems

- In contrast to other software there is no “best solution” by human translators, which can be compared with the output of the system
- I.e., for one input sentence there are many different correct translations
- Quality measurement of an MT System depends on its purposes and on the requirements of potential users.
- Possible participants in evaluation :
 - Researchers
 - Research sponsors
 - Purchasers
 - Translators

C.Vertan -ESSLI 2004 25

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

Evaluation strategies

Black Box vs. Glass Box

- MT system is seen as a black box, whose operation is treated purely in terms of its input-output behaviour
- Should not be conducted by the developers
- Tests: functionality, volume of data handled, recovery situations
- Components of the system are inspected as well as their effect in the system
- Relevant to researchers and developers
- Static analysis: checking the system without running it (automatic syntax and type checking by a compiler, manual inspection of the system, symbolic execution, data flow analysis)
- Dynamic glass box requires running the program (e.g. trying the program on many logical paths and ensuring that every logical branch is executed at least once).

C.Vertan -ESSLI 2004 26

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

Evaluation strategies

Test Suite vs. Test corpus

- Carefully constructed set of examples, each testing a particular linguistic or translation problem (e.g. different lexical and structural differences)
- Problem: it is assumed that the behaviour of a system can be projected from carefully constructed examples to real texts
- Test suite evaluations are difficult to compare
- An adequate corpus (for the domain of the system) is used as input
- Problem: it does not test systematically all possible sources of incorrect translations, but considers the most frequent constructions
- It is difficult to estimate the behaviour of the system for other types of text

C.Vertan -ESSLI 2004 27

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	----------------------	---------------	--------------------	-----------------

Evaluation - Linguistic Quality measures

- Intelligibility - measures the fluency and grammaticality of the TL text, with concern whether it faithfully conveys the meaning of the SL
- Accuracy - indicates how the translated text preserves the content of the source text. (a high intelligible sentence may not convey the meaning of the source text because of incorrect disambiguation)
- Error analysis : e.g. count the number of words inserted, modified, deleted and moved by a post-editor. However, deciding what is an acceptable translation is subjective.

C.Vertan -ESSLI 2004 28

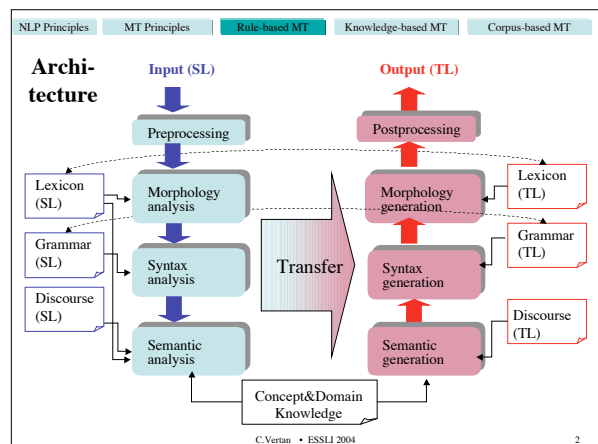
NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT



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: cri@informatik.uni-hamburg.de



NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Knowledge Sources

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

C. Vertan • ESSLI 2004 3

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Bilingual (Multilingual) Lexicon Example

Spanish Lexicon

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

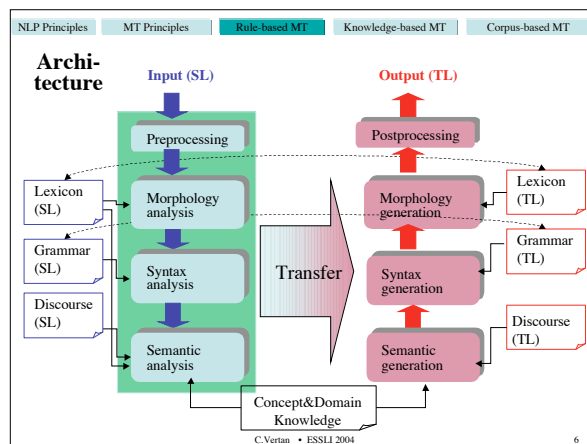
C. Vertan • ESSLI 2004 4

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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* = *Estados Unidos*
 - *Civil law* = *Código Civil*
- Sometimes abbreviations are also part of thesauri:
- E.g.:
 - *Ud(s)*=*Usted(es)* = *You (politeness)*
- Thesauri are domain specific

C.Vertan • ESSLI 2004 5



NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Morphological analysis -1-

Germ: Ich wollte die Äpfel gestern kaufen

Sp: Yo ? la ? ayer comprar

Information about:

Inflection: Äpfel (sg. masc) inflection class N23
wollen (present)

Declination: Äpfel (acusative, pl.)

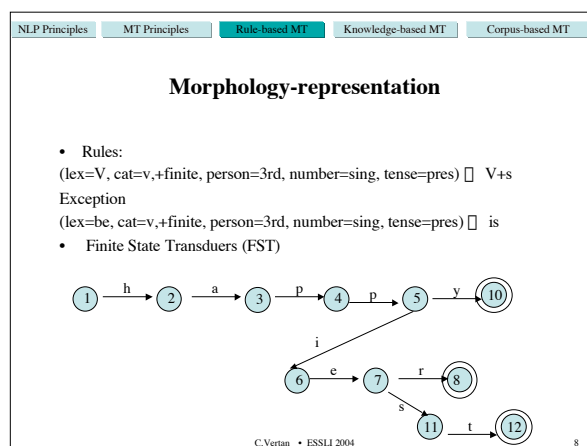
Conjugation: Ich wollte, etc.

Stem lexicon

Langenscheidt's Universal-Wörterbuch
Spanisch

Ich = Yo
die = la, las
gestern = ayer

C.Vertan • ESSLI 2004 7



NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Morphological analysis -2-

Germ: Ich wollte die Äpfel gestern kaufen

Sp: Yo hay querido la manzanas ayer comprar

Full-form lexicon

Langenscheidt's Universal-Wörterbuch

Spanisch

Ich = Yo
wollte = hay querido
die = la, las
Äpfel = manzanas
gestern = ayer

C. Verlan • ESSLI 2004 9

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Limitations of morphological analysis -1-

- From the previous example; after the morphological analysis the translation would be:
 - Yo hay querido la/las manzanas ayer comprar
- 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 Spanish order). But not all such changes can be defined by rules.

C. Verlan • ESSLI 2004 10

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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 manzana* or *las manzanas*, or complicated syntactical problems

C. Verlan • ESSLI 2004 11

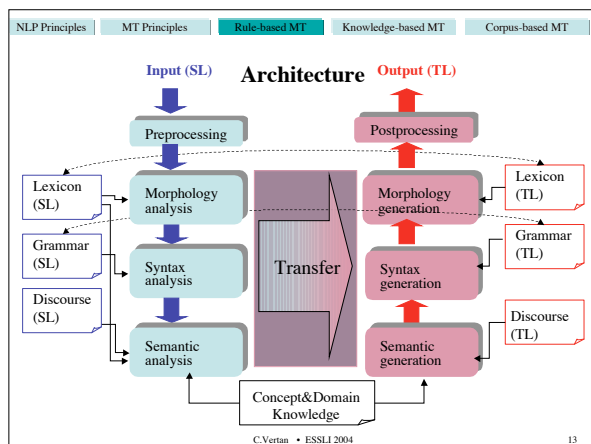
NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Syntactic and Semantic Analysis

Follow general principles in NLP

NLP Principles

C. Verlan • ESSLI 2004 12



NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Transfer levels

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

C.Vertan • ESSLI 2004 14

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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.

C.Vertan • ESSLI 2004 15

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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.) if it is outside and *pared* (sp.) 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.) → *Yo lo conozco* (sp.)
I know the solution (engl.) → *Yo sabe la solución* (sp.)

C.Vertan • ESSLI 2004 16

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Lexical transfer - Problems -2-

- Different translations depending on domain-specific/world knowledge
E.g. *biblioteca* (sp.) 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.)) which are usually not replaced in the translation process

C.Vertan • ESSLI 2004 17

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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

C.Vertan • ESSLI 2004 18

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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 re-arrangements in lexical transfer:

C.Vertan • ESSLI 2004 19

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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

C.Vertan • ESSLI 2004 20

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Syntactic Transfer - Thematic divergences-

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

C.Vertan • ESSL1 2004 21

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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

C.Vertan • ESSL1 2004 22

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Syntactic Transfer - Structural divergence

- Different sub-constituents for the same constituent.

C.Vertan • ESSL1 2004 23

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Syntactic Transfer - Lexical gaps-

- For such cases special rules must be provided

C.Vertan • ESSL1 2004 24

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Syntactic Transfer - Lexicalization-

- Languages distribute semantic content differently within a sentence

C.Vertan • ESSLI 2004 25

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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.))

C.Vertan • ESSLI 2004 26

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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.)

C.Vertan • ESSLI 2004 27

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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

C.Vertan • ESSLI 2004 28

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Semantic Transfer - Example

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

C. Verlan • ESSLI 2004 29

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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:
- $[and, VPE, [just, E]] \square [acabar, E, Subj, VPS] :- VPE \square VPS$

C. Verlan • ESSLI 2004 30

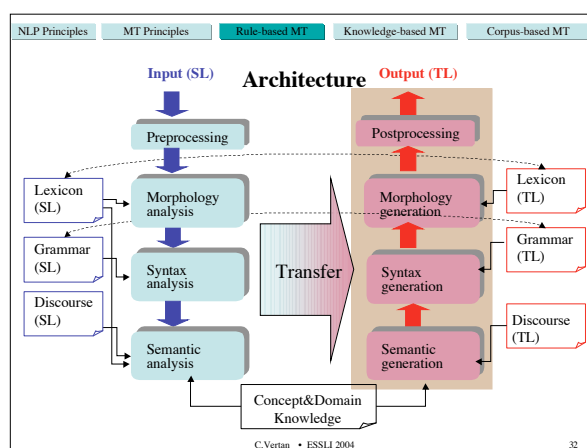
NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

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

C. Verlan • ESSLI 2004 31



NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	--------------------	-----------------

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.

C.Vertan • ESSLT 2004 33

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	--------------------	-----------------

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

C.Vertan • ESSLT 2004 34

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	--------------------	-----------------

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.

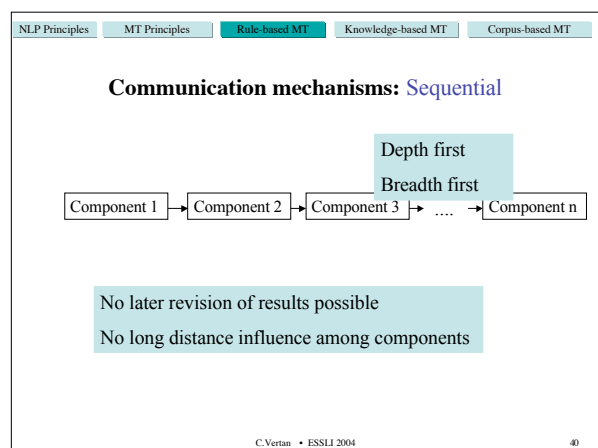
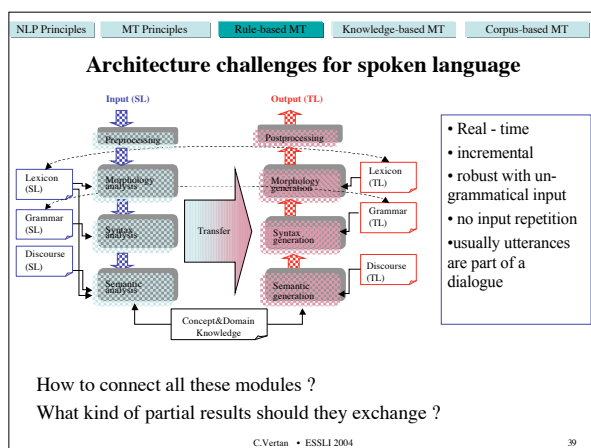
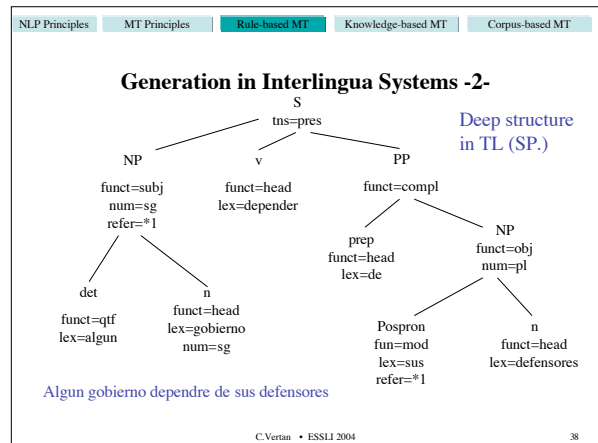
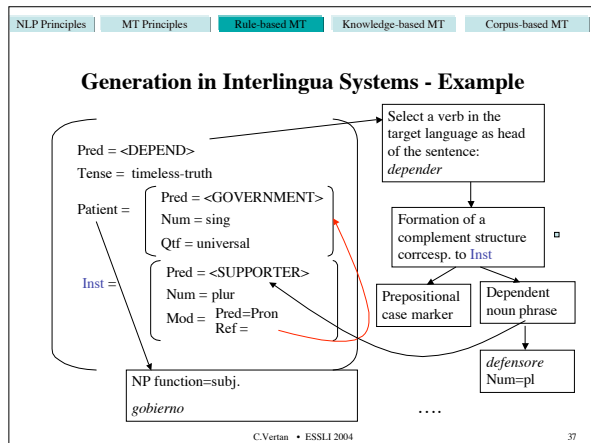
C.Vertan • ESSLT 2004 35

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	--------------------	-----------------

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)

C.Vertan • ESSLT 2004 36



NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Communication mechanisms: Cascade

Elementary interactive model
Briscoe's model

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

C. Verlan • ESSLI 2004 41

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Classical Communication mechanisms: Tree structure (sequential multi-layered)

C. Verlan • ESSLI 2004 42

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Communication mechanisms: Blackboard

HEARSAY
Verbobil II

Bottlenecks due to central scheduling

C. Verlan • ESSLI 2004 43

NLP Principles MT Principles **Rule-based MT** Knowledge-based MT Corpus-based MT

Communication mechanisms: Multi-agent

ASL
Verbobil I

Too intensive communication among components

C. Verlan • ESSLI 2004 44



Knowledge 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: cri@informatik.uni-hamburg.de

Most part of the slides from this section I owe with explicit permission from Walther v. Hahn

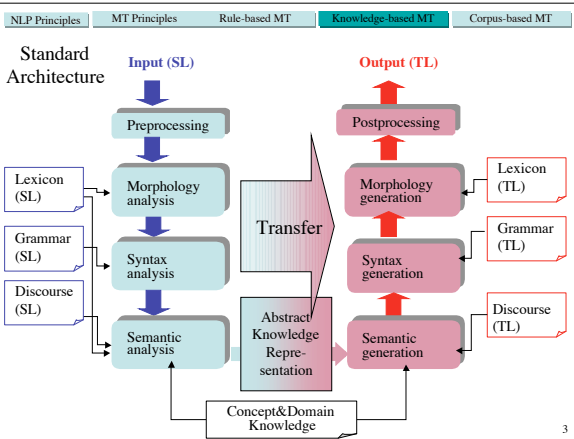
NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Outline

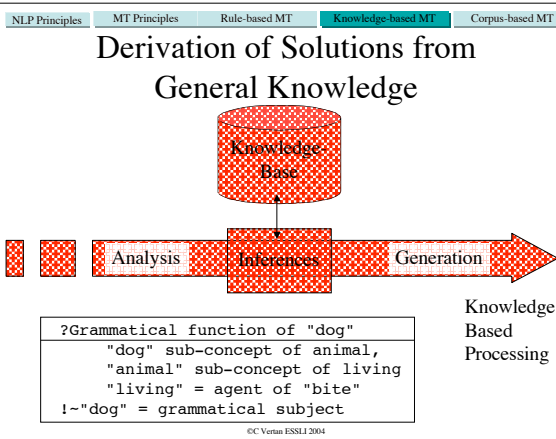
- What is knowledge?
- Why knowledge in Machine Translation?
- Linguistic examples
- 3 Examples of systems
 - KBMT 1989
 - DBR-MAT 1999
 - Verbmobil 2001
- The Semantic Web Idea

©C Vertan ESSLI 2004

2



3



©C Vertan ESSLI 2004

4

NLP Principles
MT Principles
Rule-based MT
Knowledge-based MT
Corpus-based MT

Terminological Knowledge = Knowledge-based System?

- In a sense, systems that use terminological material (in a systematic order according to the domain), can be called knowledge based systems.
- However, the ontological knowledge (conceptual ordering) of the field is not declarative, but implicit in the ordering of the terminology (nomenclature). The ontology is not visible directly.

©C Veritas ESSL1 2004
5

NLP Principles
MT Principles
Rule-based MT
Knowledge-based MT
Corpus-based MT

"Real" Knowledge in MT

©C Veritas ESSL1 2004
6

NLP Principles
MT Principles
Rule-based MT
Knowledge-based MT
Corpus-based MT

Structural Transfer with Interlingua

Tasks:

- Content analysis from the Lexicon, Morphology, Syntax, Semantics, Pragmatics
- = Mapping of the Input onto:
 - 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

©C Veritas ESSL1 2004
7

NLP Principles
MT Principles
Rule-based MT
Knowledge-based MT
Corpus-based MT

Interlingua Expression

```

[ request (referent (_5747))
  presuppose (exists (_4340)) ]
some (_4340)
[ unique (_4407)
  single (_4407)
  instance (_4407, person)
  propval (Person, _4407, sex, male)
  [ some (_4725)
    [ unique (_5033)
      single (_5033)
      instance (_5033, project)
      propval (project, _5033, name, str
        (LOKI)) ]
    instance (_4725, leading)
    propval (leading, _4725, theta, _5033)
    propval (leading, _4725, alpha, _4407)
    topic(_4407) ] ] ]

```

Interlingua systems
construct a deep
meaning representation
by using additional
knowledge of some sort,
at least about what
sentences are about.

Presupposition part
of an
Interlingua expression

©C Veritas ESSL1 2004
8

NLP Principles
MT Principles
Rule-based MT
Knowledge-based MT
Corpus-based MT

Knowledge Layers

Knowledge used in an MT system may be

- Conceptual knowledge (“ontology”, “upper model”)
- World knowledge (chemical laws, e.g.)
- Factual knowledge (situational knowledge) about the actual state of affairs

- Some examples:

©C Veritas ESSL1 2004
9

NLP Principles
MT Principles
Rule-based MT
Knowledge-based MT
Corpus-based MT

Conceptual Knowledge

$\Box b \text{ Bear } (b) \Box \text{ Animal } (b)$ *“Bears are animals”*

$\Box b \text{ Bear } (b) \Box \text{ SpeciesOf } (b) = \text{Ursidae}$
“Bears belong to the species of Ursidae”

$\Box x,y \text{ part of } (x,y) \ \& \ \text{PhysicalThing } (y) \Box \text{ PhysicalThing } (x)$
“Parts of physical objects are physical objects”

©C Veritas ESSL1 2004
10

NLP Principles
MT Principles
Rule-based MT
Knowledge-based MT
Corpus-based MT

World Knowledge

$\Box x \text{ PhysThing } (x) \Box \Box s \text{ size}(x) = s$
“All physical things have a size”

Bear (Pooh) *“Pooh is a bear”*

T (Area (Poland, SqMiles (233000)), AD1426)
“In 1426 Poland had the size of 239.000 square meters”

©C Veritas ESSL1 2004
11

NLP Principles
MT Principles
Rule-based MT
Knowledge-based MT
Corpus-based MT

Situation Knowledge

In (Pooh, LivingRoom3)
 “Pooh is in the living room”

Female (Speaker23)
 “The current speaker is female”

(most examples from Russell/Norvig)

©C Veritas ESSL1 2004
12

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	---------------------------	-----------------

Do We Need Knowledge in MT at all?

- Whatever a MT system does, it needs at least some of the before mentioned knowledge layers. If the functionality of the system is not very demanding, some parts of the conceptual knowledge can be described in the lexicon, e.g. under semantic features, sub-categorization, roles or constraints.
- But any system will run into difficulties ...

©C Veritas ESSLI 2004 13

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	---------------------------	-----------------

Disambiguation

source: situation knowledge

"I will **go** to my office in an hour"

- "**Voy** dentro de una hora en mi despacho"
- "**Viajo** dentro de una hora en mi despacho"
- "**Volo** dentro de una hora en mi despacho"

(engl. □ esp.)

dependent on *how far away* my office is.

Similarly, in Romanian **a merge** □ either **walk** or **drive**

©C Veritas ESSLI 2004 14

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	---------------------------	-----------------

Anaphora antecedents

source: concepts/facts

"She took the ice cream from the fridge and ate **it**"

- "Ella ha tomando la tarta del refrigerador y **la** ha comido" (not "**el**")

(engl. □ esp.)

©C Veritas ESSLI 2004 15

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	---------------------------	-----------------

Pronouns

source: facts

"**They** go"

- "**Ți** merg" / "**Ele** merg" / "**Ei Ți ele** merg"

(engl. □ rom.)

Dependent on whether the set of persons ("they") contains men or woman or both

©C Veritas ESSLI 2004 16

NLP Principles MT Principles Rule-based MT **Knowledge-based MT** Corpus-based MT

Lexical Specification

source: concept hierarchy/facts

“cousin” □ “cousin” or “cousine”
(engl. □ fr.)

“Uncle” □ “farbror” or “morbror”
(germ. □ dan.)

©C Veritas ESSLI 2004 17

NLP Principles MT Principles Rule-based MT **Knowledge-based MT** Corpus-based MT

3 Examples of the use of knowledge in MT

- KBMT 1989
- DBR-MAT 1999
- Verbmobil 2001

©C Veritas ESSLI 2004 18

NLP Principles MT Principles Rule-based MT **Knowledge-based MT** Corpus-based MT

Example I: The **KBMT** project

Assumptions behind KBMT:

One “functionally complete” meaning representation can serve for translations to a number of languages, no total representation of human understanding of a text is necessary for machine translation, applicable to relatively unambiguous, e.g. technical documents.

©C Veritas ESSLI 2004 19

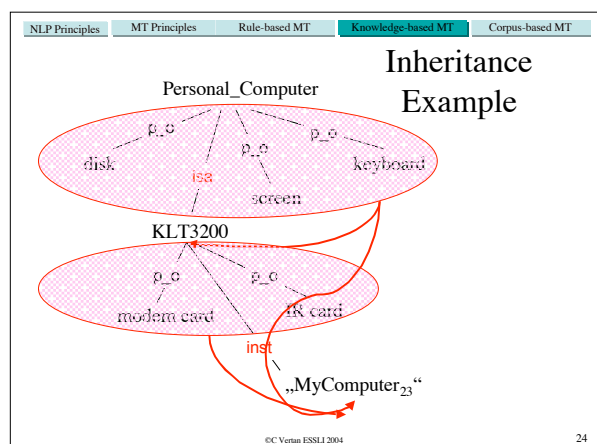
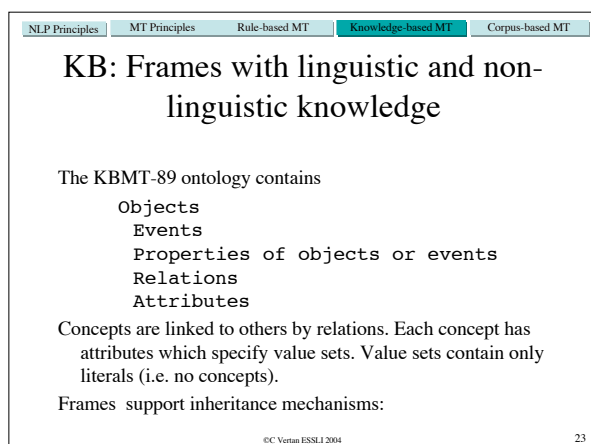
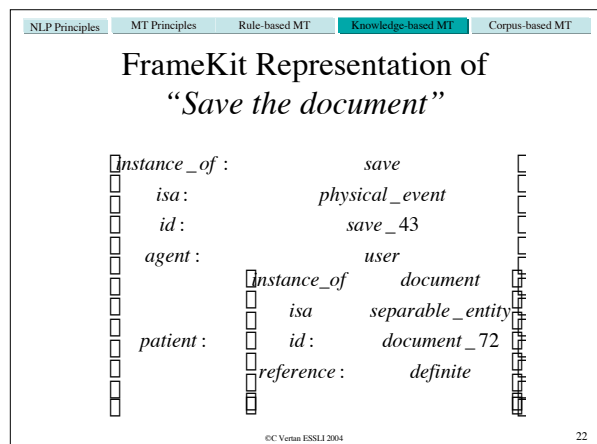
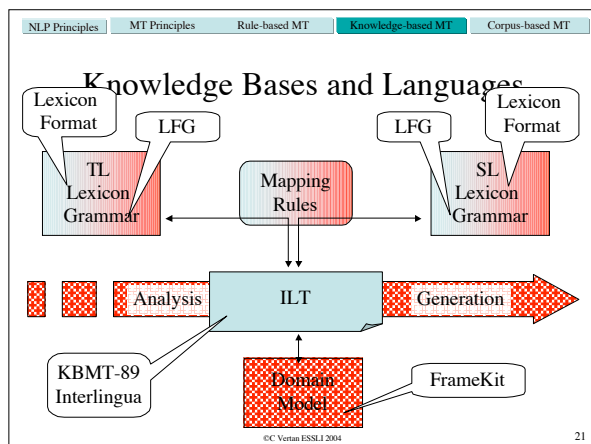
NLP Principles MT Principles Rule-based MT **Knowledge-based MT** Corpus-based MT

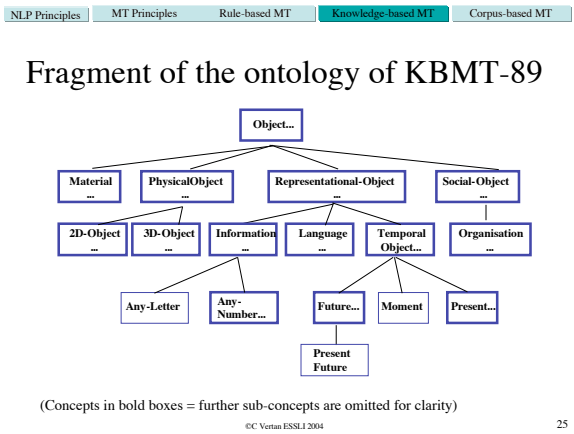
Basic Components of a KBMT System

An ontology of concepts (“domain model”, “ontology”)
Source language (SL) lexicon and grammar for the analysis process
Target language (TL) lexicon and grammar for the generation processes
Mapping rules between the Interlingua and SL/TL syntax.

Technical aim of KBMT-89: Bidirectional translations for single sentences of English and Japanese in the domain of PC manuals.

©C Veritas ESSLI 2004 20





NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

ILT for “Get the diagnostics diskette from the back of this manual”

- [*RECEIVE
(AGENT *READER)
(THEME [*DISKETTE
(NUMBER SINGULAR)
(PURPOSE [*DIAGNOSE
(NUMBER SINGULAR)])
(REFERENCE DEFINITE)])
(SOURCE [*BACK-OF-3D
(NUMBER SINGULAR)
(REFERENCE DEFINITE)
(PART-OF [*MANUAL
(NUMBER SINGULAR)
(REFERENCE DEFINITE)
(NEAR +)])])
(TENSE PRESENT)
(MOOD IMPERATIVE)
(CLAUSAL-MARK +)
(NUMBER-BULLET [*ANY-NUMBER
(CARDINALITY 1)])]

©C Verlan ESSLI 2004 26

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

Example II: The Verbmobil System

Transfer system for English, Japanese and German
Spontaneous speech input, on-line processing near real-time.
Sentence representation: VIT (Verbmobil Interface Term) for necessary linguistic and non-linguistic information.
no representation of semantic contents or pragmatic sense in VITs

- VITs represent **dialogue acts** (→ syntactic disambiguation, repair of gaps in speech hypotheses, and
- prosody** as indicator for structural boundaries, particle interpretation and sentence mood.

©C Verlan ESSLI 2004 27

NLP Principles MT Principles Rule-based MT Knowledge-based MT Corpus-based MT

```

VIT: vltID (VERBMOBIL ID) 10, 808
[ word(1, 2, 1),
  word(coming, 3, 1127),
  word(at, 4, 1136),
  word(the, 5, 1128),
  word(beginning, 6, 1135),
  word(of, 7, 1135),
  word("August", 8, 1134) ] ] ]
index(1138, 1125, 135), %Index
[beginning(1125, 137), %Conditions
arg3(1135, 137, 138),
come(1127, 135),
arg3(1127, 135, 136),
decl(1137, h43),
pron(1126, 136),
at(1136, 135, 137),
mov(1114, 138, 888),
def(1128, 137, h42, h41),
udef(1131, 138, h45, h44) ],
[ in_g(1126, 1125), in_g(1137, 1138), %Constraints
in_g(1127, 1125), in_g(1128, 1130),
in_g(1121, 1123), in_g(1134, 1132),
in_g(1135, 1129), in_g(1136, 1125),
leq(1125, h41), leq(1125, h43),
leq(1128, h42), leq(1129, h44),
leq(1130, h43), leq(1132, h45),
leq(1133, h43) ],
[ s_sort(135, situation), %Sorts
s_sort(137, time),
s_sort(138, time),
[ dialog_act(1125, inform), %Discourse
dir(1136, no),
prontype1136, third, std() ],
[ cas(136, none), %Syntax
gend(136, masc),
num(136, sg), num(137, sg), num(138, sg),
pcase(1135, 139, c4),
[ ta_aspect(135, progr), %Tense and Aspect
ta_mood(135, ind),
ta_perf(135, nonperf),
ta_tense(135, pres), ]
[ prose_accent(1135, progr) ], %Prosody
]

```

©C Verlan ESSLI 2004 28

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Multilingual Semantic Data Base

Base form of words,
Lexical semantic decomposition,
Interlingual representation, if available,
Semantic class (nominals, quantifiers, verbs, modifiers, etc.),
Syntactic valency with mappings onto grammatical functions and thematic roles (linking).

Ontological sorts, e.g.:

- abstract
 - property, field, info-content, institution, symbol
- space-time
 - temporal
 - situation
 - meeting_sit, communicate_sit, action_sit, ...)
 - time
 - entity
 - object
 - agentive, thing
 - location

Selectional restrictions on arguments.

©C Veritas ESSLI 2004

29

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Dialog Act Structure

©C Veritas ESSLI 2004

30

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Example III: The DBR-MAT System

Functionality:

Technical translators are supported.

- They can ask conceptual questions concerning,
 - intensional definitions,
 - extensional definitions,
 - characteristics and attributes,
 - hyponyms or meronyms
 - Examples
 - etc.
- ask for pictures,
- recursively ask about the explanations.

©C Veritas ESSLI 2004

31

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

The DBR-MAT Paradigm

©C Veritas ESSLI 2004

32

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Conceptual Graphs

Examples of complex conceptual relations ("situations")

[SITUATION: [OIL FRAGMENT: {*}] (IN)]

[WATER: {*}]

(CHAR) [PHYSICAL STATE: disj{MEMBRANE, DROPS, COLLOID, EMULSION, SOLUTION}].

[SITUATION:

[WASTE WATER: {*}] (CONTAIN)]

[OIL FRAGMENT: {*}]

(ATTR) [FLOATING]

(ATTR) [ROUGHLY DISPERSED]

(PTNT) [PRECIPITATION]].

©C Veritas ESSLJ 2004

33

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Connecting words and concepts

©C Veritas ESSLJ 2004

34

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Multiple inheritance for differences in conceptualizing

©C Veritas ESSLJ 2004

35

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Knowledge Base

The knowledge base of DBR-MAT contains the following objects:

KB Objects	Example
Concepts	[OIL SEPARATOR]
Individuals	the separator C334
Conceptual relations	(part_of)
Contexts (situations)	[SITUATION: [WASTE WATER: {*}] -> ... [SEPARATOR]
A type hierarchy	[OIL SEPARATOR]

©C Veritas ESSLJ 2004

36

9

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
Traversing rules				
Submenu	Item	Evaluated Conceptual Relations	Inheritance	
What is?		Types of... · All + ATTR, Char, PART_OF	✓	
Types of	All	Superconcepts + subconcepts + sister concepts		
	General	All superconcepts from the hierarchy		
	Concrete	All subconcepts from the hierarchy		
	Similar	All sister concepts from the hierarchy		
Characteristics	All	Attributes + Who + Object + How + Where	✓	
	Attributes	ATTR + CHAR	✓	
	Who	AGNT		
	Object	OBJ + PTNT		
	How	INST		
	Where	LOC + DEST + FROM + IN + TO		
More...		All remaining relations		
Examples		Individual concepts		
Want All		All mentioned above, without duplicates	✓	

©C Veritas ESSL1 2004

37

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
Critical problems of knowledge-based systems are still				
1. The huge effort to build up knowledge bases, 2. A practical definition of the size ("coverage") of the knowledge base, and 3. The choice of the representation language and its necessary logical/formal properties.				



©C Veritas ESSL1 2004

38

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
Advantage of Knowledge Bases				
• Using knowledge bases the developer definitely knows, what is represented where, although he cannot predict, what can be derived with the inference rules. With implicit and procedural (local) representations there is no method to check multiple representation. • Declarative knowledge sources, are global, can be maintained in isolation, can be exchanged and may be used in other inference machines or grammars. It even can be used in other systems than translation systems.				



©C Veritas ESSL1 2004

39

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
The Semantic Web Idea				
• This basic idea of declarative and modularized knowledge has become very important since the famous paper of Berners-Lee in 2001 on the "semantic web". • The Semantic Web (according to the definition of the W3C consortium) is • "the abstract representation of data on the World Wide Web, based on the RDF standards and other standards to be defined. It is being developed by the W3C, in collaboration with a large number of researchers and industrial partners."				

©C Veritas ESSL1 2004

40

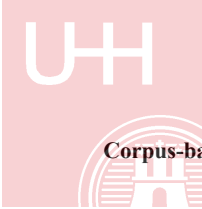
NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	---------------------------	-----------------

The Semantic Web Idea

- *The semantic web will serve as one (or several) ontology(ies) to which all WWW objects refer and which can be used consequently for web operations like data mining, information extraction, summarization, etc"*
- **and translation!**
- The first obvious result of the semantic web activities for translation is, that widely accepted ontologies of specific domains can be used as knowledge bases for machine translation. This solves point (1) to (3) of the above mentioned list of problems.

©C Vortan ESSLI 2004 41

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**



Corpus-based Maschine translation

Cristina Vertan
 University of Hamburg • Informatics Department
 Natural Language Systems Group

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

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Contents

- General Principles
- Statistical Machine Translation
- Example-based Machine Translation
- Corpus based MT vs. Rule-based MT.



©C. Vertan - ESSLI 2004 2

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

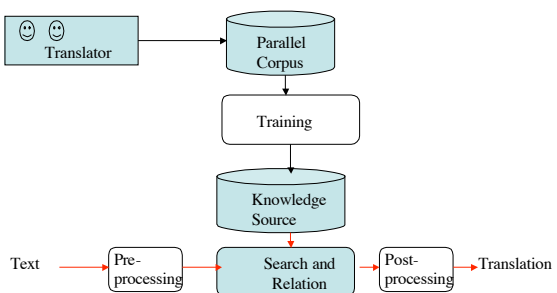
General Principles

- The linguistic phenomena in both languages as well as the transfer rules are no longer linguistically described but derived automatically from a parallel corpus.
- First an aligned corpus is built
- Next step is a training phase, in which are calculated the connections between elements in the source language as well as in the target language (sometimes the results are called „knowledge sources“).
- The translation is the result of 2 processes:
 - A search process (of elements in the source language)
 - A best-evaluated relation with a target expression
- There are 2 types of corpus-based MT systems
 - Example based MT - The translation of a source text is based on translation examples in the database
 - Statistical MT - the alignment information from the corpus is used for the training of a statistical translation model

©C. Vertan - ESSLI 2004 3

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Generical Architecture of a corpus-based MT



```

graph TD
    Translator[Translator] --> PC[(Parallel Corpus)]
    PC --> Training[Training]
    Training --> KS[(Knowledge Source)]
    KS --> SR[Search and Relation]
    Text[Text] --> PP[Pre-processing]
    PP --> SR
    SR --> PostPP[Post-processing]
    PostPP --> Translation[Translation]
  
```

©C. Vertan - ESSLI 2004 4

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Aligned Corpus

- A parallel Corpus:
 - Is a collection of texts in at least 2 languages. It is extremely important that the content is the same for both texts.
 - Examples: Official Documents from EU , Newspapers in Countries with more than 1 official language
 - Contains markers (tags) for content-identical elements (Sentences, Paragraphs) in Texts
- The parallel aligned Corpus has to be adequate for the translation domain.
- When searching such corpora the main problem is, that 1 chunk in the source has more than 1 translation in the target language and the choice is made according to the context.

©C. Verlan - ESSLI 2004

5

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Parallel aligned Corpus - Example 1-

<DOC de-news-1996-10-02-1>
 <H1>
 Streit um baden-wuerttembergischen
 Nachtragshaushalt
 </H1>
 Die Oppositionsfaktionen im baden
 wuerttembergischen Landtag haben scharfe
 Kritik an der Finanzpolitik der CDU/FDP.
 Koalition geuebt. Bei der Vorlage des zweiten
 Nachtragshaushalts fuer das laufende
 Haushaltsjahr begruesseten sie zwar die strikte
 Begrenzung der Neuverschuldung , gespart werde
 aber vor allem bei den Familien und damit am
 falschen Fleck.Finanzminister Mayer-Vorfelder
 verteidigte den eingeschlagenen Sparkurs. Der
 Nachtragshaushalt soll rund 1,1 Mrd. DM
 Deckungsluecke ausgleichen, die vor allem durch
 Steuerausfaelle im Haushalt klaffen. Rund 800
 Mio. DM sollen durch Investitions- und
 Sachmittelkuerzungen erbracht werden.
 Einsparungen im Personalbereich werden mit 130
 Mio. DM beziffert.

<DOC de-news-1996-10-02-1>
 <H1>
 Baden-Wuerttemberg supplementary budget
 dispute
 </H1>
 The opposition parties in Baden-Wuerttemberg's
 Landtag have strongly criticized the financial
 policies of the governing CDU/FDP coalition. Upon
 presentation of the second supplementary budget for
 the current budget year, they approved of the strict
 limitation of new borrowing, but said that savings
 were going to be realized in the wrong place - on the
 backs of families. Finance Minister Mayer Vorfelder
 defended the budget, saying it would equal out a
 shortfall of 1.1 billion marks in state finances, which
 was caused primarily by tax losses. Cuts to
 investments and materials are expected to yield 800
 million marks. Savings on personnel are estimated
 at 130 million Marks.

<DOC de-news-1996-10-02-2>

<DOC de-news-1996-10-02-2>

Paragraph-Alignment

©C. Verlan - ESSLI 2004

6

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Parallel aligned corpus - Example 2-

<DOC de-news-1996-10-02-1>
 Streit um baden - wuerttembergischen
 Nachtragshaushalt

 Die Oppositionsfaktionen im baden -
 wuerttembergischen Landtag haben scharfe Kritik an
 der Finanzpolitik der CDU / FDP - Koalition geuebt .

 Bei der Vorlage des zweiten Nachtragshaushalts fuer
 das laufende Haushaltsjahr begruesseten sie zwar die
 strikte Begrenzung der Neuverschuldung , gespart
 werde aber vor allem bei den Familien und damit am
 falschen Fleck .

 Finanzminister Mayer - Vorfelder verteidigte den
 eingeschlagenen Sparkurs . Der Nachtragshaushalt soll
 rund 1,1 Mrd. DM Deckungsluecke ausgleichen , die
 vor allem durch Steuerausfaelle im Haushalt klaffen .

<DOC de-news-1996-10-02-1>
 Baden - Wurttemberg supplementary budget dispute

 The opposition parties in Baden - Wurttemberg 's
 Landtag have strongly criticized the financial policies
 of the governing CDU / FDP coalition .

 Upon presentation of the second supplementary
 budget for the current budget year , they approved of
 the strict limitation of new borrowing , but said that
 savings were going to be realized in the wrong place -
 - on the backs of families .

 Finance Minister Mayer - Vorfelder defended the
 budget , saying it would equal out a shortfall of 1.1
 billion marks in state finances , which was caused
 primarily by tax losses .
 ...

Sentence-Alignment

2 Sentences in the DE-
Corpus correspond to 1
Sentence in the EN-
Corpus

©C. Verlan - ESSLI 2004

7

NLP Principles

MT Principles

Rule-based MT

Knowledge-based MT

Corpus-based MT

Parallel aligned Corpus - Example 3 -

Die ₁ Oppositionsfaktionen ₂ im ₃ baden - wuerttembergischen ₄ Landtag ₅ haben
 scharfe ₆ Kritik ₇ an der Finanzpolitik ₈ der ₉ CDU / FDP - ₁₀ Koalition ₁₁ geuebt ₁₂ .

The (1) opposition parties (2) in (3) Baden - Wurttemberg 's(4) Landtag(5) have strongly (6)
 criticized(7+12) the financial policies(8) of (9) the governing CDU / FDP (10) coalition(11)
 .

Fertility of a source word = the number of words in the
 target text

 e.g. $fertility(Oppositionsfaktionen) = 2$

 Distortion = Source and target words do not appear in the
 same place e.g. Koalition und coalition

©C. Verlan - ESSLI 2004

8

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Alignment-Methods

- **Manual :**
 - Extreme time consuming, because for real applications the corpus has to be really big.
 - Specialists with very good knowledge in both languages are needed
- Automatic with help of statistical procedures
 - E.g. length-based methods (number of words in the source and target text has to be close one to another)
 - Difficult to identify at the word-level, because for e.g. in:

Haben Scharfe Kritik geübt □ have strongly criticized

The POS change and the semantic combinations are different.

©C. Verlan - ESSLI 2004 9

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

- General principles
- Statistical machine Translation
- Example-based Machine Translation
- Corpus-based MT vs. Example-based MT

←

©C. Verlan - ESSLI 2004 10

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Statistical MT-Principles - 1 -

- Given:
 - A source sentence (e.g. in German.): $D = d_1, \dots, d_n$ (d_i are the words) which has to be translated into a sentence (in English for e.g.) $E = e_1, \dots, e_m$.
 - A parallel aligned German-English corpus

Between all translation possibilities it is searched the one with the highest probability.

- This means mathematically :

$$\hat{e}^m = \arg \max_{e^m} \{ \Pr(e^m | d^n) \}$$
- Depending on how this probability is calculated there are different models for the translation.

©C. Verlan - ESSLI 2004 11

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Statistical MT -Model 1

- Das Source-Channel Modell (used very often):
 - Following decomposition is used:

$$\Pr(e^m | d^n) = \Pr(e^m) \Pr(d^n | e^m)$$

Language model - give the probability that e^m is a correct English sentence

Translation model
Give the probability that in the corpus a sentence e^m will be found which is the translation of d^n

Both models are dependent of parameters, which are calculated in the training phase

©C. Verlan - ESSLI 2004 12

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Statistical MT- Model 2-

- Direct Maximum Entropy Translation Model
 - The original probability is calculated directly, following different translation features (mathematically is a function with parameters)

$$\Pr(e^m | d^n)$$
- Alignment Model
 - A new parameter is introduced, which models the alignment-mapping. Here features like Fertility and Distortion are considered

©C. Verlan - ESSLI 2004 13

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Advantages of Statistical MT

- Use no linguistic knowledge (as long as the alignment of the corpus is done automatically)
- Loose dependencies between constituents can be modelled better with statistical models as with rules
- It is especially indicated to be used in embedded systems e.g. in Speech Systems, where a language model already is defined (for the speech recognizer)

©C. Verlan - ESSLI 2004 14

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Well-known problems with Statistical MT

- New field, there are few systems which can be evaluated. (Verbmobil, Translation of Canadian parliament debates)
- Exceptions can be trained difficult
- Morphology:
 - Inflected forms of the same word are treated as not-related words. E.g. the Word *diriger* in French is translated with *führen* or *leiten* in German. For each one of the 39 inflected forms of the word the model has to be trained (which is translated with *führen* and which with *leiten* muss).
- Not-local dependencies are difficult to be trained. The System produces usually correct word-translations but in an incorrect order
- Probabilities for rare words are not to be trusted.
- The models are very sensible to data-changes.

©C. Verlan - ESSLI 2004 15

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Example of incorrect Translations with statistical MT -1-

- Source text: Permettez que je donne un exemple à la chambre.
- Correct translation: Let me give the House an example
- System-Translation: Let me give an example in the House

The model gives the highest probability for the word à the word in.

© Manning undSchütz S. 489
©C. Verlan - ESSLI 2004 16

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Examples of incorrect translations with statistical MT -2-

- Source text: Vous avez besoin **de toute l'aide** disponible
- Correct Translation: You need all the help you can get
- System-Translation: You need **of the whole benefits** available.

In the aligned corpus the translation of **toute l'aide** is found with **whole benefits** and simply used further

© Manning undSchütz S. 489

©C. Verlan - ESSLI 2004 17

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

- General principles
- Statistical MT
- **Example-based Machine Translation - EBMT** ←
- Corpus-based vs. Rule-based MT

©C. Verlan - ESSLI 2004 18

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

General Principles of EBMT

- A parallel Corpus is useec
- Part of the input text are compared with source chunks in the corpus
- The translation of the founded parts are put together and form the translation.

©C. Verlan - ESSLI 2004 19

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Translation pyramid for EBMT

©C. Verlan - ESSLI 2004 @ A. Way 20

NLP Principles MT Principles Rule-based MT Knowledge-based MT **Corpus-based MT**

EBMT - Example

- input: John went to the baker's
- The corpus contains:
 - John went to school - Jan ging in die Schule
 - the baker's is the whole day opened. - Die Bäckerei ist ganztags geöffnet
- Identified
 - John went to - Jan ging in
 - The baker's - die Bäckerei
- The translation:
 - Jan ging in die Bäckerei.

©C. Verlan - ESSLI 2004 21

NLP Principles MT Principles Rule-based MT Knowledge-based MT **Corpus-based MT**

Corpus-Tagging for EBMT

- It is possible to mark in the Corpus Words of morphemes, which delimit a clear co-text Quantifikatorc, Konjunktions, Pronouns, Question markers, etc.
- E.g. <QUANT> all uses
- <QUANT> tous usages

©C. Verlan - ESSLI 2004 22

NLP Principles MT Principles Rule-based MT Knowledge-based MT **Corpus-based MT**

Length and Size of Examples

- The size of the example database varies between some hundreds and 800.000 Sentences.
- The bigger is the Database the better works the system
- There is no ideal length for the examples:
 - The longer the examples, the lower the chance for a match
 - The shorter the example the bigger is the chance to have some ambiguities
- Usually the standard unit for the examples is a sentence

©C. Verlan - ESSLI 2004 23

NLP Principles MT Principles Rule-based MT Knowledge-based MT **Corpus-based MT**

EBMT with linguistic knowledge

- The Translation patterns are not words, but syntactical structures in both languages with corresponding links

©C. Verlan - ESSLI 2004 24

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Input for Matching

- The problem is to find what parts of the input can be retrieved in the database
- This is done through a combination of string-based and statistical-based methods (e.g. big probability for multi-word lexemes).

©C. Verlan - ESSLI 2004 25

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

What are Translation Memories ?

- Translation Memories TM are big databases with translation examples
- They contain also a search mechanism.
- Example: TRADOS - a TM-System for all old 12 EU-languages.

©C. Verlan - ESSLI 2004 26

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

EBMT and Translation Memories

- The search mechanism (analysis) is the same .
- TM have no combination phase.
- Translation Memories are help for translators but do not replace them, therefore the output is a list with translation alternatives for parts or complete sentences.

©C. Verlan - ESSLI 2004 27

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

EBMT and Statistical MT

- There are trials to combine the two approaches
- The idea is exceptions which cannot be statistical modelled or very often expressions to be filtered through example search.
- E.g.. in Verbmobil System:
 - Ich denke mal - I think
 - Also wir wollten - well, we plan to
 - In unserer Abteilung ein neues Netzwerk aufbauen - set up a new network in our department.

©C. Verlan - ESSLI 2004 28

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

- General principles
- Statistical MT
- Example-based Machine Translation - EBMT
- **Corpus-based vs. Rule-based MT**



©C. Verlan - ESSLI 2004 29

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Comparison of linguistic and empirical methods - 1-

- In Verbmobil-System (German-English-Japanese Speech-to-Speech System) were 5 MT approaches implemented, 1 transfer-bases, 1 statistical and 1 example based.
- After the evaluation of the mistakes:
 - Semantic Transfer 62 %
 - Example based MT 35%
 - Statistical MT 29%

©C. Verlan - ESSLI 2004 30

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Comparison of linguistic and empirical methods - 2-

- Most problems of the empirical approaches are due to:
 - Word orde: the target language model is not trained accordingly
 - Disambiguation: Very difficult for prepositions which are translated according to the context
 - No partial translation: statistical translation process only sentences. If the speech recognizer contains mistakes or „false starts“ the quality of the translation decrease rapidly.
 - Problems with verb paricles: in this case is a morphological preprocessing needed, but this increses very much the processing time

©C. Verlan - ESSLI 2004 31

NLP Principles | MT Principles | Rule-based MT | Knowledge-based MT | **Corpus-based MT**

Comparison of lingusitic and empirical methods - Examples 1-

- Spoken: Okay. Dann treffen wir uns um acht Uhr sechs auf dem Hauptbahnhof in Hamburg. Ist das okay?
- Recognized: okay dann treffen wir uns um acht Uhr sechs auf dem Hauptbahnhof in hamburg. Ist das okay
- Transfer: okay | that is okay
- Example: and then we will meet in Hamburg at eight six is that okay
- Statistical: okay, then we will meet at eight oh six at the main train station, in hamburg, is that okay.

©C. Verlan - ESSLI 2004 32

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	--------------------	-----------------

Comparison of linguistic and empirical methods - Examples 2-

- Spoken: Dann muss ich Ihnen gleich mal sagen, die ersten zwei Oktoberwochen sieht's bei mir ganz schlecht aus.
- Recognized: dann muss ich Ihnen gleich mal sagen die erste zwei Oktober-Wochen sieht's bei mir ganz schlecht aus sind
- Transfer: -
- Example: *then I have to tell you right away it looks quite bad on my end in on the first two weeks of October I*
- Statistical: *then I have to tell you right away, the first two weeks of October is very bad for me, is*

©C. Verlan - ESSLI 2004 33

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	--------------------	-----------------

Comparison of linguistic and empirical methods - Examples 3-

- Spoken: Das tut mir leid, da habe ich - , muss ich auf eine Messe.
- Recognized: das tut mir leid da ich mus ich eine Messe
- Transfer: I'm sorry about that I must a fair there
- Example: -
- Statistical: I am sorry, I have got a fair.

©C. Verlan - ESSLI 2004 34

NLP Principles	MT Principles	Rule-based MT	Knowledge-based MT	Corpus-based MT
----------------	---------------	---------------	--------------------	-----------------

Corpus-Linguistics and Statistics for Linguists Links

- http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/content_s.htm - Course on Corpus Linguistics
- www.coli.uni-sb.de/~christer/stat_cl.ps Linguists guide to statistics.

©C. Verlan - ESSLI 2004 35

Literature

- *"An Introduction to Machine Translation "*, J. Hutchins and H. Somers, Academic Press, 1992
- *"Machine Translation - an Introductory Guide "*, D. Arnold,L. Balkan, NCC Blackwell London, 1994 (
<http://clwww.essex.ac.uk/MTbook/PostScript/>)
- *"A Survey of current Paradigms in Machine Translation "*, B. J. Dorr, P. W. Jordan, J. W. Benoir, 2000 (
<http://www.umiacs.umd.edu/~bonnie/>)
- *"Übersetzungswissenschaft. Eine Neuorientierung "*, Mary Snell-Hornby(Hrsg.), Francke Verlag Tuebingen, 1986
- *"Linguistic and Computational Techniques in Machine Translation System Design "*, P.Whitelock, K. Kilby, UCL Press, 1995
- *"Example-based Machine Translation"*, A. Way, m. Carl, Kluwer Academic Press 2003