EUROMATRIX:
Statistical and hybrid machine translation between all European languages

A Project funded by the European Community under the Sixth Framework Programme (IST-5-034291-STP)

Hybrid Architectures for Machine Translation

Andreas Eisele, Saarland University & DFKI GmbH

EuroMatrix 2nd Machine Translation Marathon, Wandlitz, May 14, 2008

Information Society
Technologies

Hybrid Architectures for Machine Translation

Structure of Presentation

- Motivation
- A menagerie of hybrid architectures
- What we (and others) did so far and what could be done
- Conclusion and next steps

Motivation I

Different approaches to MT have complementary PROs and CONs:

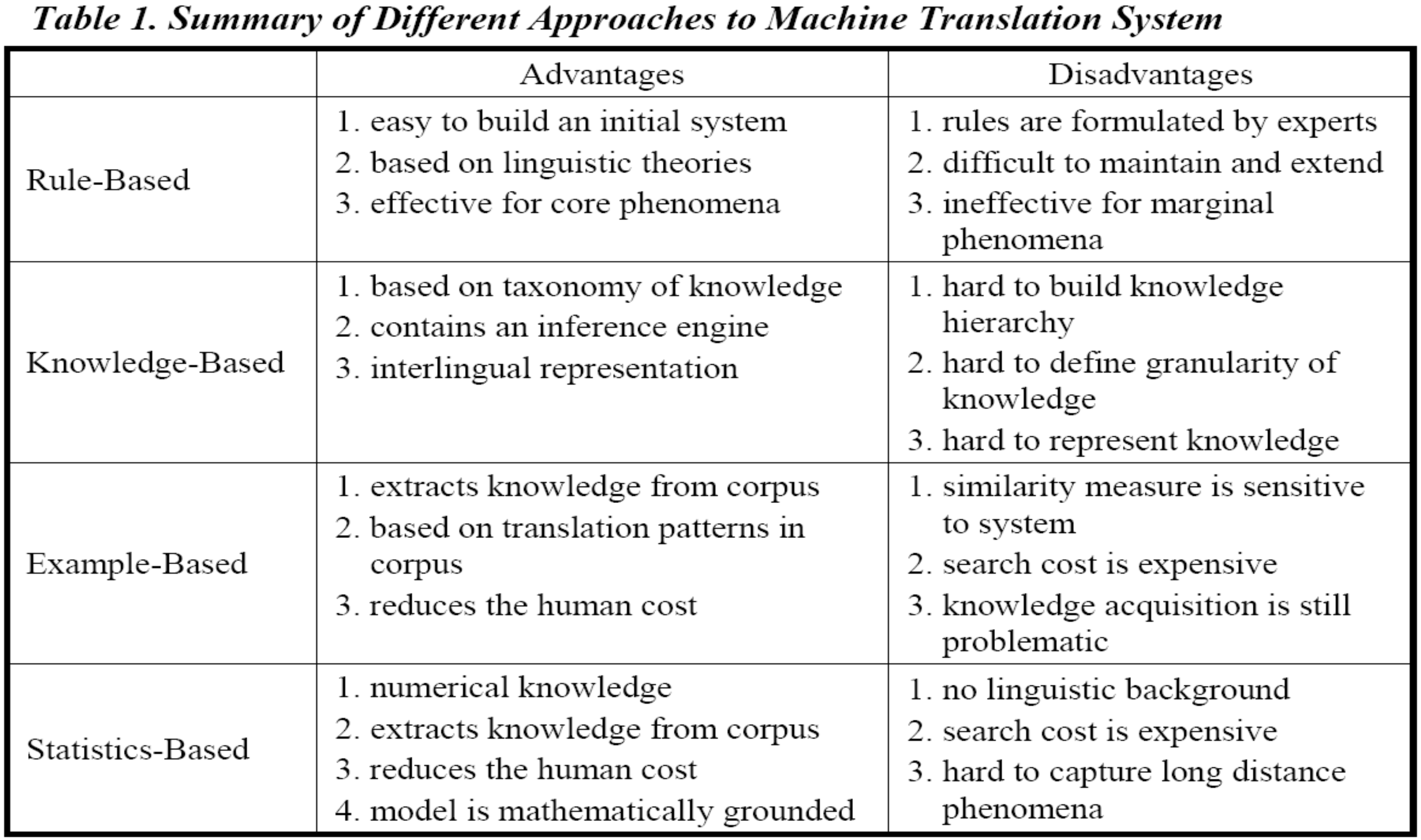


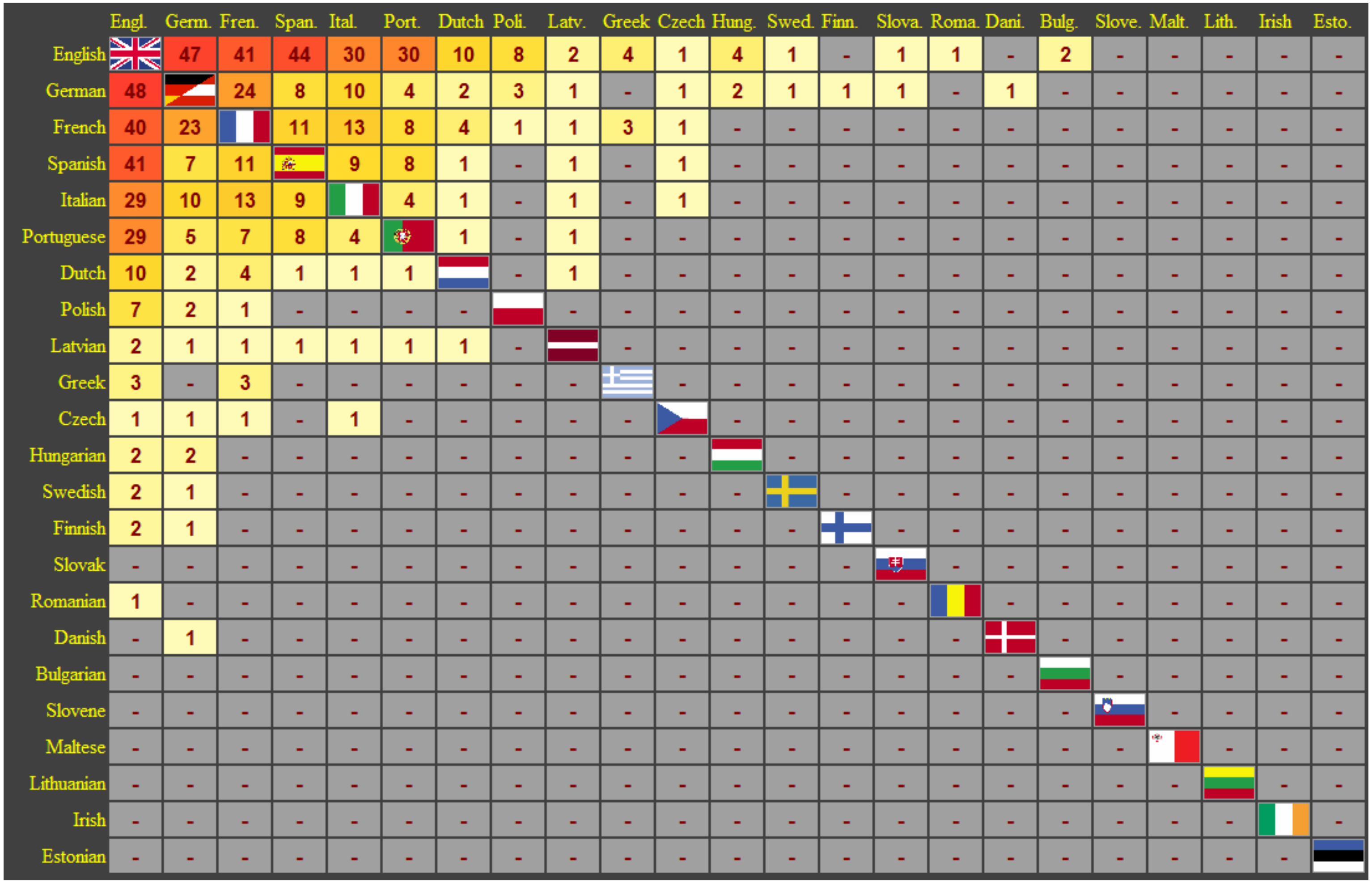
Table 1. Summary of Different Approaches to Machine Translation System

	Advantages	Disadvantages
Rule-Based	1. easy to build an initial system 2. based on linguistic theories 3. effective for core phenomena	1. rules are formulated by experts 2. difficult to maintain and extend 3. ineffective for marginal phenomena
Knowledge-Based	1. based on taxonomy of knowledge 2. contains an inference engine 3. interlingual representation	1. hard to build knowledge hierarchy 2. hard to define granularity of knowledge 3. hard to represent knowledge
Example-Based	1. extracts knowledge from corpus 2. based on translation patterns in corpus 3. reduces the human cost	1. similarity measure is sensitive to system 2. search cost is expensive 3. knowledge acquisition is still problematic
Statistics-Based	1. numerical knowledge 2. extracts knowledge from corpus 3. reduces the human cost 4. model is mathematically grounded	1. no linguistic background 2. search cost is expensive 3. hard to capture long distance phenomena

Source: Chen & Chen: A Hybrid Approach to Machine Translation System Design, Computational Linguistics and Chinese Language Processing, 1996

Motivation II

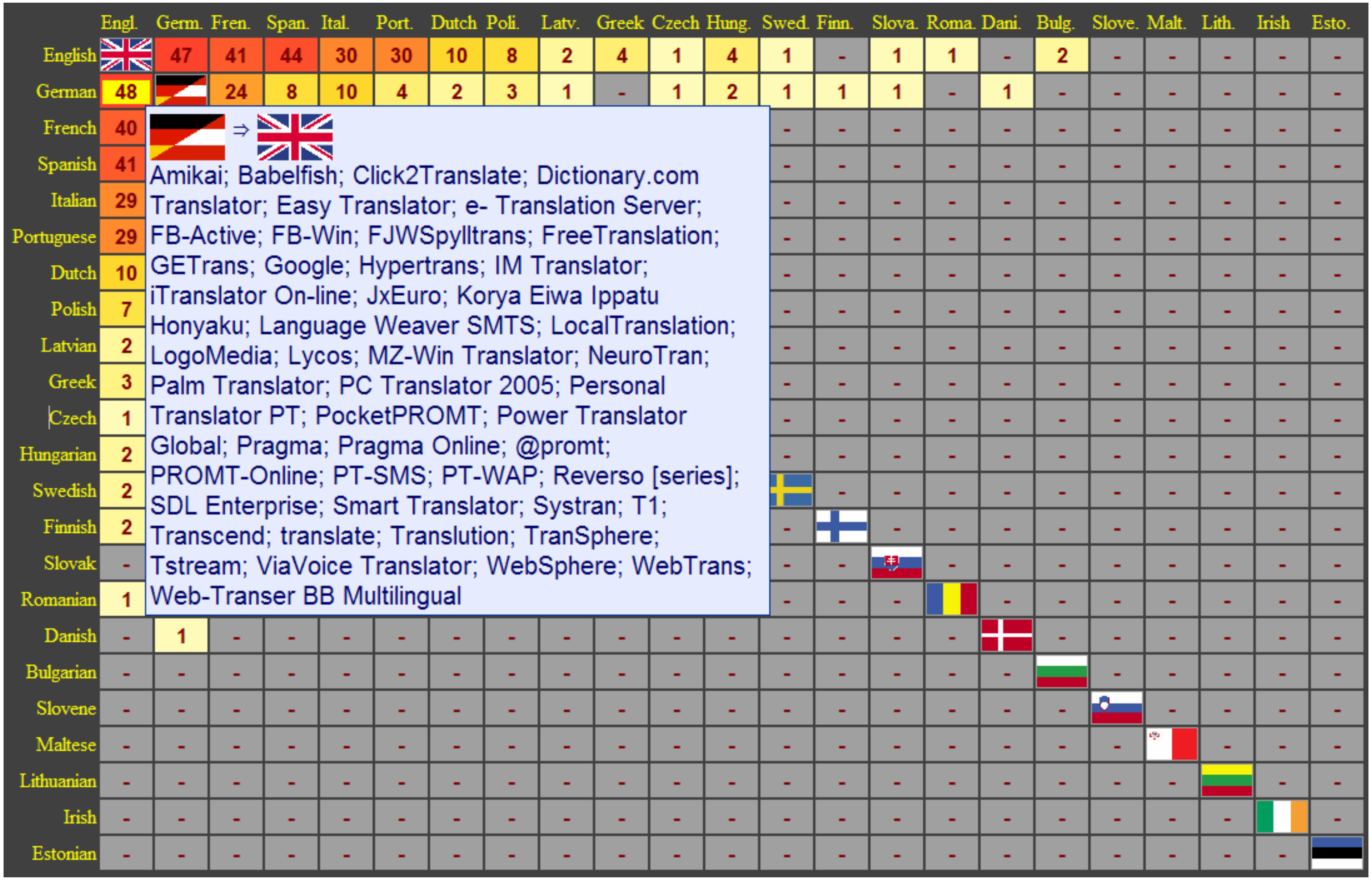
MT systems per language pair [according to Hutchins 2005]



	Engl.	Germ.	Fren.	Span.	Ital.	Port.	Dutch	Poli.	Latv.	Greek	Czech	Hung.	Swed.	Finn.	Slova.	Roma.	Dani.	Bulg.	Slove.	Malt.	Lith.	Irish	Esto.
English	🇬🇧	47	41	44	30	30	10	8	2	4	1	4	1	-	1	1	-	2	-	-	-	-	-
German	48	🇩🇪	24	8	10	4	2	3	1	-	1	2	1	1	1	-	1	-	-	-	-	-	-
French	40	23	🇫🇷	11	13	8	4	1	1	3	1	-	-	-	-	-	-	-	-	-	-	-	-
Spanish	41	7	11	🇪🇸	9	8	1	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Italian	29	10	13	9	🇮🇹	4	1	-	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-
Portuguese	29	5	7	8	4	🇵🇹	1	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dutch	10	2	4	1	1	1	🇳🇱	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Polish	7	2	1	-	-	-	-	🇵🇱	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Latvian	2	1	1	1	1	1	1	-	🇱🇻	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Greek	3	-	3	-	-	-	-	-	-	🇬🇷	-	-	-	-	-	-	-	-	-	-	-	-	-
Czech	1	1	1	-	1	-	-	-	-	-	🇨🇿	-	-	-	-	-	-	-	-	-	-	-	-
Hungarian	2	2	-	-	-	-	-	-	-	-	-	🇭🇺	-	-	-	-	-	-	-	-	-	-	-
Swedish	2	1	-	-	-	-	-	-	-	-	-	-	🇸🇪	-	-	-	-	-	-	-	-	-	-
Finnish	2	1	-	-	-	-	-	-	-	-	-	-	-	🇫🇮	-	-	-	-	-	-	-	-	-
Slovak	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🇸🇰	-	-	-	-	-	-	-	-
Romanian	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🇷🇴	-	-	-	-	-	-	-
Danish	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🇩🇰	-	-	-	-	-	-
Bulgarian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🇧🇬	-	-	-	-	-
Slovene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🇸🇮	-	-	-	-
Maltese	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🇲🇹	-	-	-
Lithuanian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🇱🇹	-	-
Irish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🇮🇪	-
Estonian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🇪🇪

Motivation II

MT systems per language pair [according to Hutchins 2005]

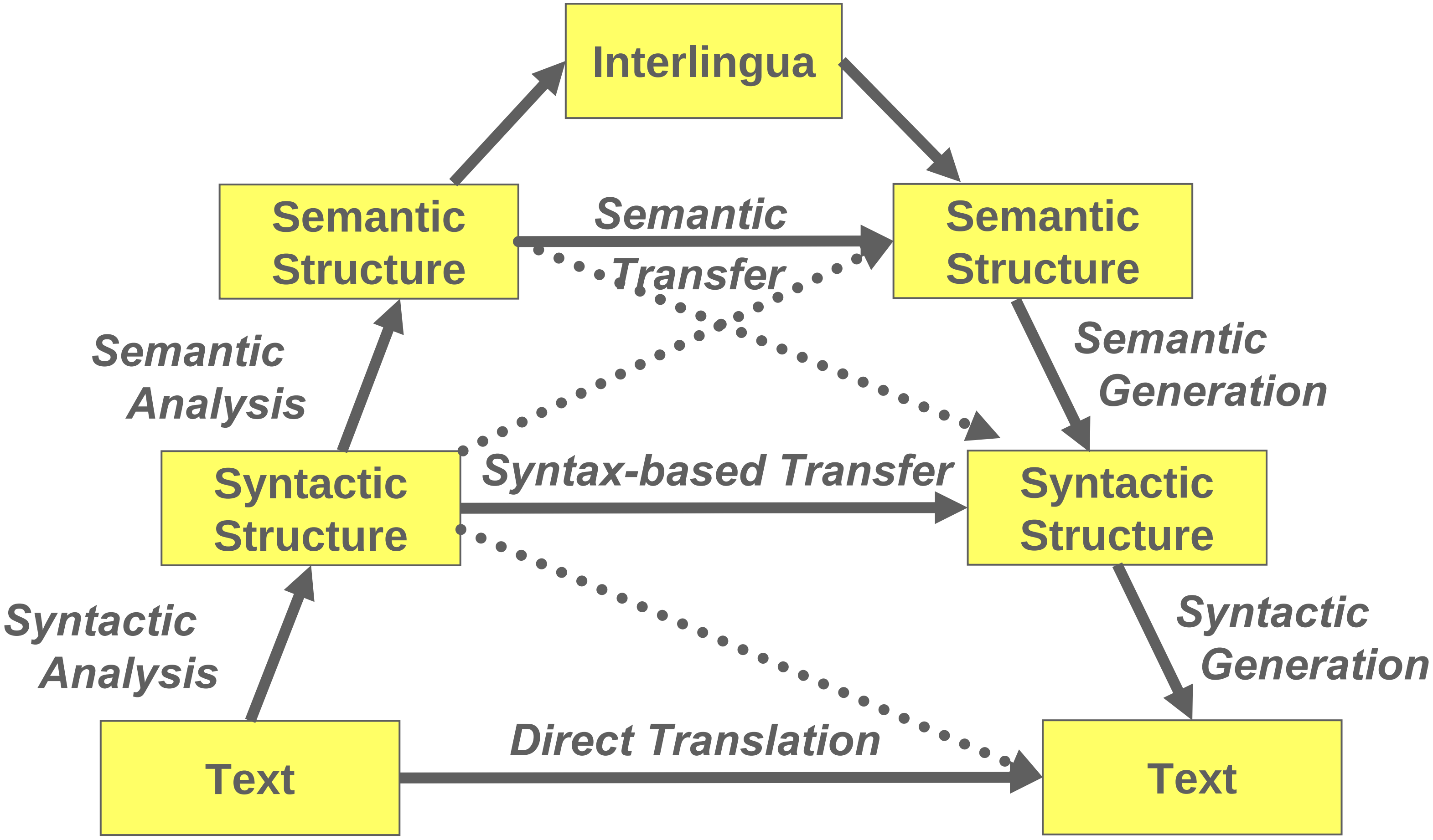


	Engl.	Germ.	Fren.	Span.	Ital.	Port.	Dutch	Poli.	Latv.	Greek	Czech	Hung.	Swed.	Finn.	Slova.	Roma.	Dani.	Bulg.	Slove.	Malt.	Lith.	Irish	Esto.
English		47	41	44	30	30	10	8	2	4	1	4	1	-	1	1	-	2	-	-	-	-	-
German	48		24	8	10	4	2	3	1	-	1	2	1	1	1	-	1	-	-	-	-	-	-
French	40												-	-	-	-	-	-	-	-	-	-	-
Spanish	41												-	-	-	-	-	-	-	-	-	-	-
Italian	29												-	-	-	-	-	-	-	-	-	-	-
Portuguese	29												-	-	-	-	-	-	-	-	-	-	-
Dutch	10												-	-	-	-	-	-	-	-	-	-	-
Polish	7												-	-	-	-	-	-	-	-	-	-	-
Latvian	2												-	-	-	-	-	-	-	-	-	-	-
Greek	3												-	-	-	-	-	-	-	-	-	-	-
Czech	1												-	-	-	-	-	-	-	-	-	-	-
Hungarian	2												-	-	-	-	-	-	-	-	-	-	-
Swedish	2													-	-	-	-	-	-	-	-	-	-
Finnish	2												-		-	-	-	-	-	-	-	-	-
Slovak	-												-	-		-	-	-	-	-	-	-	-
Romanian	1												-	-	-		-	-	-	-	-	-	-
Danish	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-	-
Bulgarian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-	-
Slovene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-	-
Maltese	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-	-
Lithuanian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-	-
Irish	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-		-
Estonian	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	

⇒

Amikai; Babelfish; Click2Translate; Dictionary.com Translator; Easy Translator; e- Translation Server; FB-Active; FB-Win; FJWSpylltrans; FreeTranslation; GETrans; Google; Hypertrans; IM Translator; iTranslator On-line; JxEuro; Korya Eiwa Ippatu Honyaku; Language Weaver SMTS; LocalTranslation; LogoMedia; Lycos; MZ-Win Translator; NeuroTran; Palm Translator; PC Translator 2005; Personal Translator PT; PocketPROMT; Power Translator Global; Pragma; Pragma Online; @promt; PROMT-Online; PT-SMS; PT-WAP; Reverso [series]; SDL Enterprise; Smart Translator; Systran; T1; Transcend; translate; Translution; TranSphere; Tstream; ViaVoice Translator; WebSphere; WebTrans; Web-Transer BB Multilingual

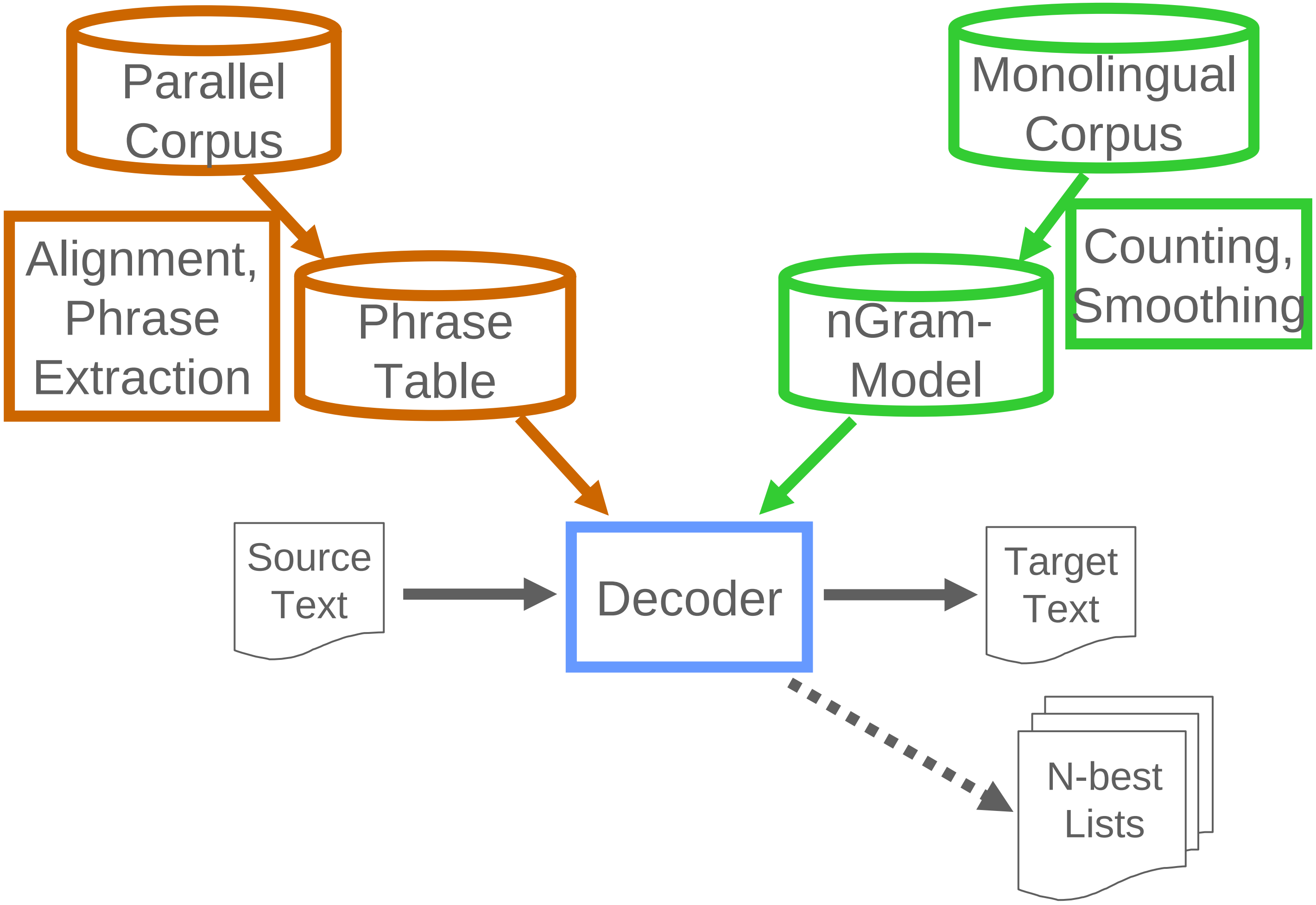
Schematic overview of RMBT architectures



The „Vauquois-Triangle“

Schematic overview of SMT architecture

Relevant knowledge is extracted automatically from text



Strengths and Weaknesses of SMT vs. RMBT

(RBMT:translate pro ⟷ SMT:Koehn 2005, examples from EuroParl)

EN: ***I wish the negotiators continued success with their work in this important area.***

RBMT: ***Ich wünsche, dass die Unterhändler Erfolg mit ihrer Arbeit in diesem wichtigen Bereich fortsetzten.***

continued: Verb instead of adjective

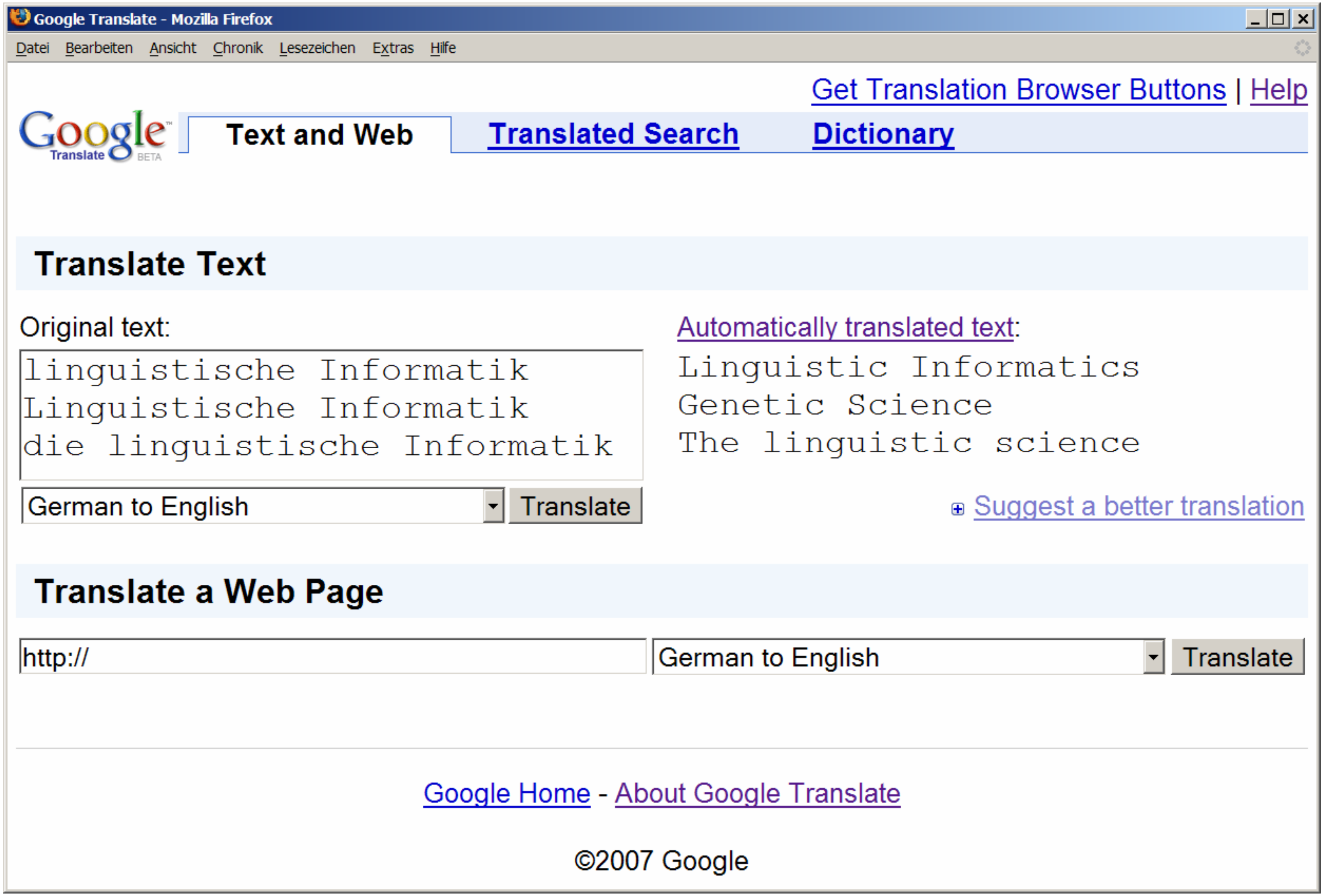
SMT: ***Ich wünsche der Verhandlungsführer fortgesetzte Erfolg bei ihrer Arbeit in diesem wichtigen Bereich.***

three wrong inflectional endings

Strengths and Weaknesses of SMT vs. RMBT

Englisch	RMBT: translate pro	SMT: Koehn 2005
We seem sometimes to have lost sight of this fact.	*Wir scheinen manchmal Anblick dieser Tatsache verloren zu haben.*	*Manchmal scheinen wir aus den Augen verloren haben, diese Tatsache.*
The leaders of Europe have not formulated a clear vision.	*Die Leiter von Europa haben keine klare Vision formuliert.*	*Die Führung Europas nicht formuliert eine klare Vision.*
I would like to close with a procedural motion.	*Ich möchte mit einer verfahrenstechnischen Bewegung schließen.*	*Ich möchte abschließend eine Frage zur Geschäftsordnung ε.*

Problems with Reliability of Lexicon Acquisition


Google Translate - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
Get Translation Browser Buttons | Help
Google Translate BETA
Text and Web
Translated Search
Dictionary

Translate Text

Original text:

```
linguistische Informatik
Linguistische Informatik
die linguistische Informatik
```

German to English | Translate

Automatically translated text:

```
Linguistic Informatics
Genetic Science
The linguistic science
```

Suggest a better translation

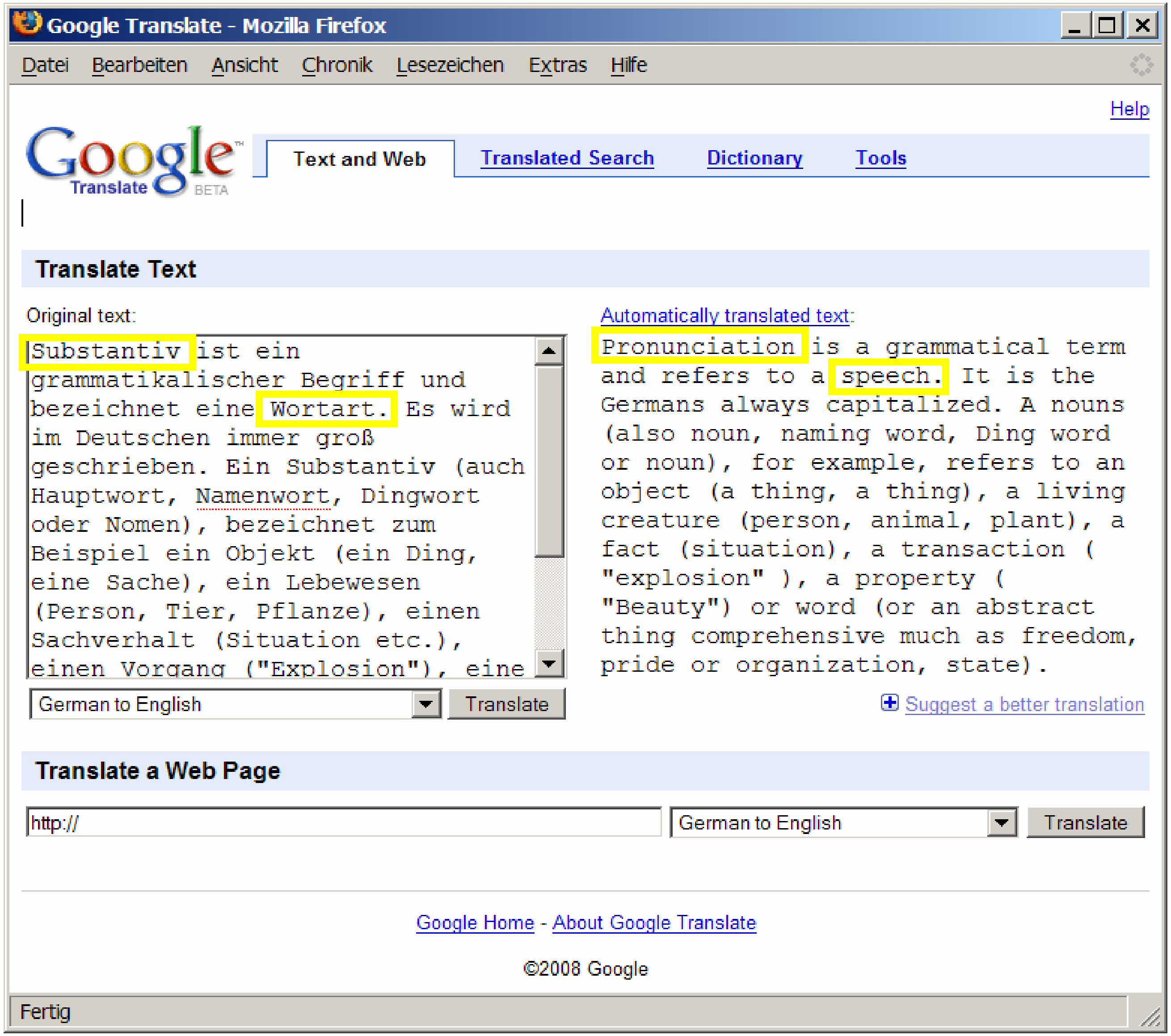
Translate a Web Page

http:// | German to English | Translate

Google Home - About Google Translate
©2007 Google

[November 2007, corrected in the meantime]

More Examples of Reliability Problems



Google Translate - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
Help
Google Translate BETA
Text and Web
Translated Search
Dictionary
Tools

Translate Text

Original text:

Substantiv ist ein grammatikalischer Begriff und bezeichnet eine Wortart. Es wird im Deutschen immer groß geschrieben. Ein Substantiv (auch Hauptwort, Namenwort, Dingwort oder Nomen), bezeichnet zum Beispiel ein Objekt (ein Ding, eine Sache), ein Lebewesen (Person, Tier, Pflanze), einen Sachverhalt (Situation etc.), einen Vorgang ("Explosion"), eine

German to English ▾ Translate

Automatically translated text:

Pronunciation is a grammatical term and refers to a speech. It is the Germans always capitalized. A nouns (also noun, naming word, Ding word or noun), for example, refers to an object (a thing, a thing), a living creature (person, animal, plant), a fact (situation), a transaction ("explosion"), a property ("Beauty") or word (or an abstract thing comprehensive much as freedom, pride or organization, state).

Translate a Web Page

http:// | German to English ▾ | Translate

[January 2008, corrected in the meantime]

Motivation for Hybrid Approaches to MT

In the early 90s, SMT and RBMT were seen in sharp contrast.

But advantages and disadvantages are complementary.

➔ Search for integrated methods is now seen as natural extension for both approaches

	RBMT	SMT
Syntax, Morphology	++	--
Structural Semantics	+	--
Lexical Semantics	-	+
Lexical Adaptivity	--	+
Lexical Reliability	+	-

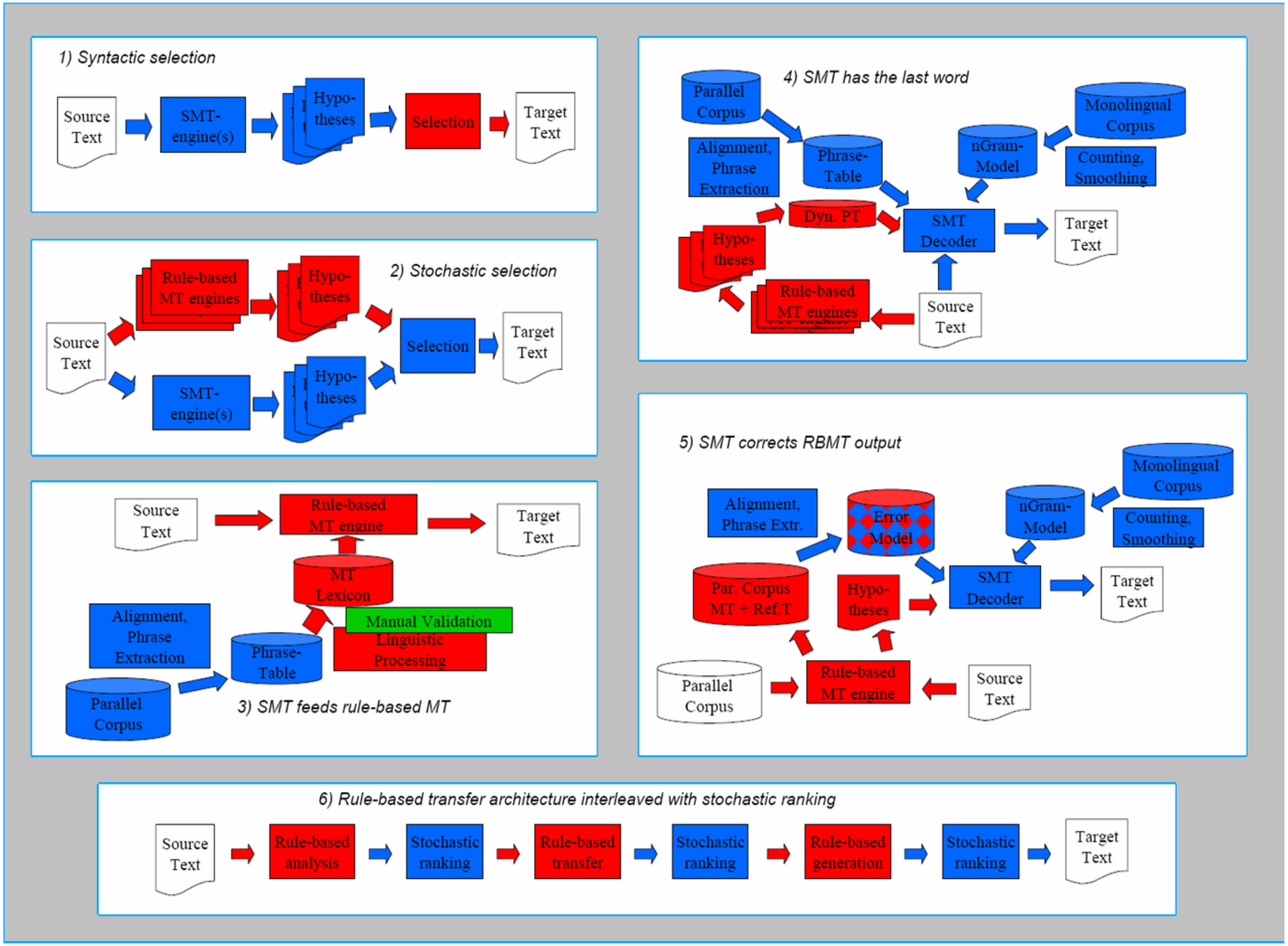
Basic Assumptions behind EuroMatrix‘ WP6

- Different MT engines tend to make different types of errors
- Combining outputs of several MT engines can improve overall quality
- This requires us to identify and combine good parts within competing candidates
- Even more improvements may be made by combining the different knowledge sources/modules in a hybrid architecture

Overview of possible hybrid architectures

From poster at WMT07:

🟦 = SMT Module
🟥 = RBMT Module



1) Syntactic selection

2) Stochastic selection

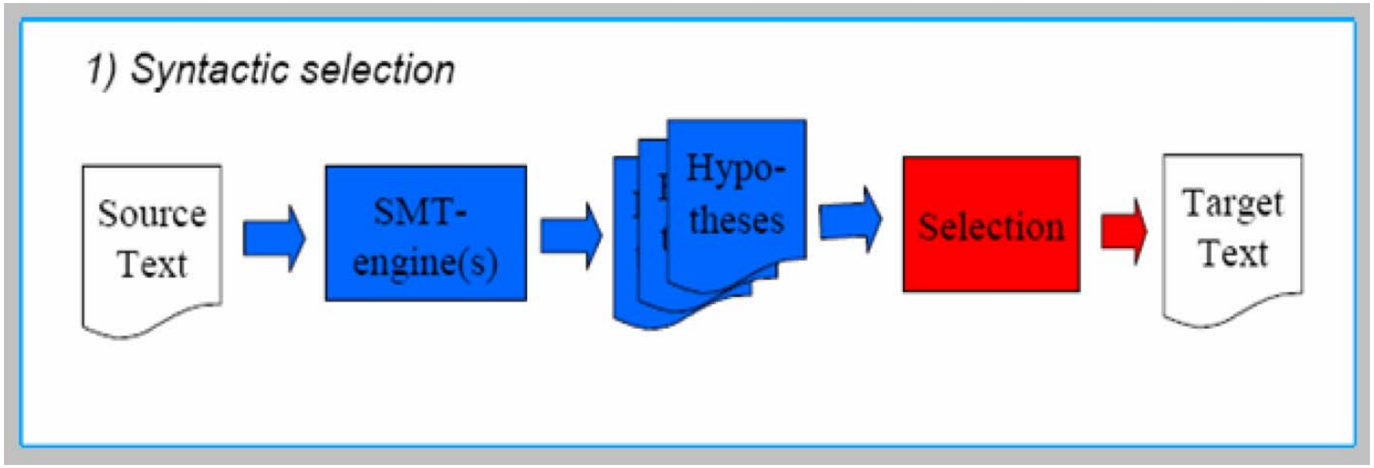
3) SMT feeds rule-based MT

4) SMT has the last word

5) SMT corrects RBMT output

6) Rule-based transfer architecture interleaved with stochastic ranking

Syntactic Selection



Motivation: SMT output often syntactically ill-formed

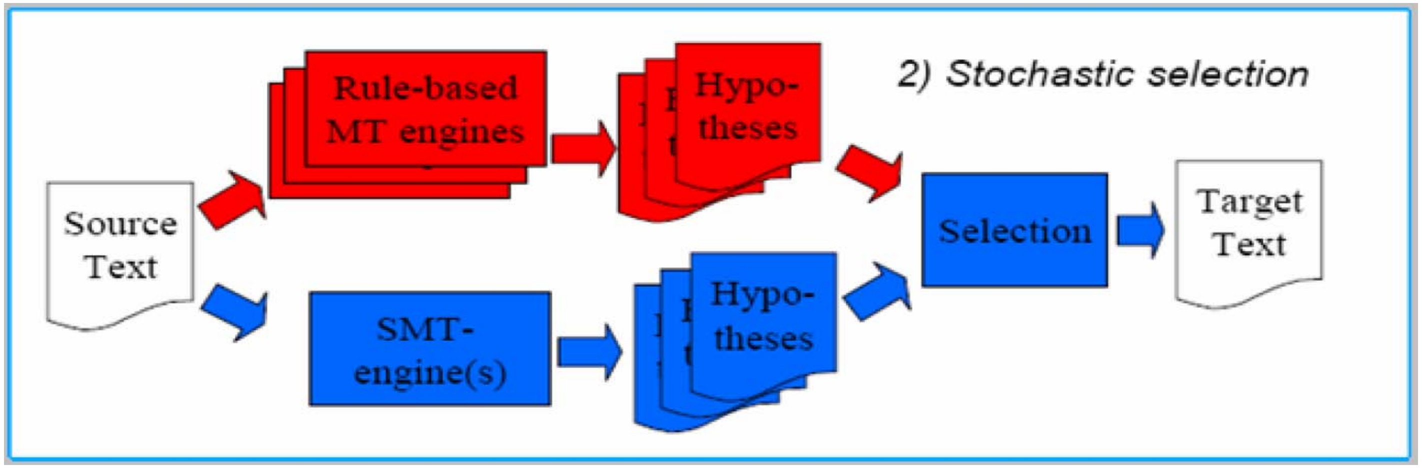
➔ Selection mechanism in SMT „generate and test“ should be enriched with syntactic knowledge

BUT:

- syntactic parsers not (yet) robust enough
- High computational cost of processing many ill-formed candidates

➔ Need to explore cues for syntactic well-formedness without full parsing

Stochastic Selection



Motivation: Selection from an increased number of candidates can improve overall quality

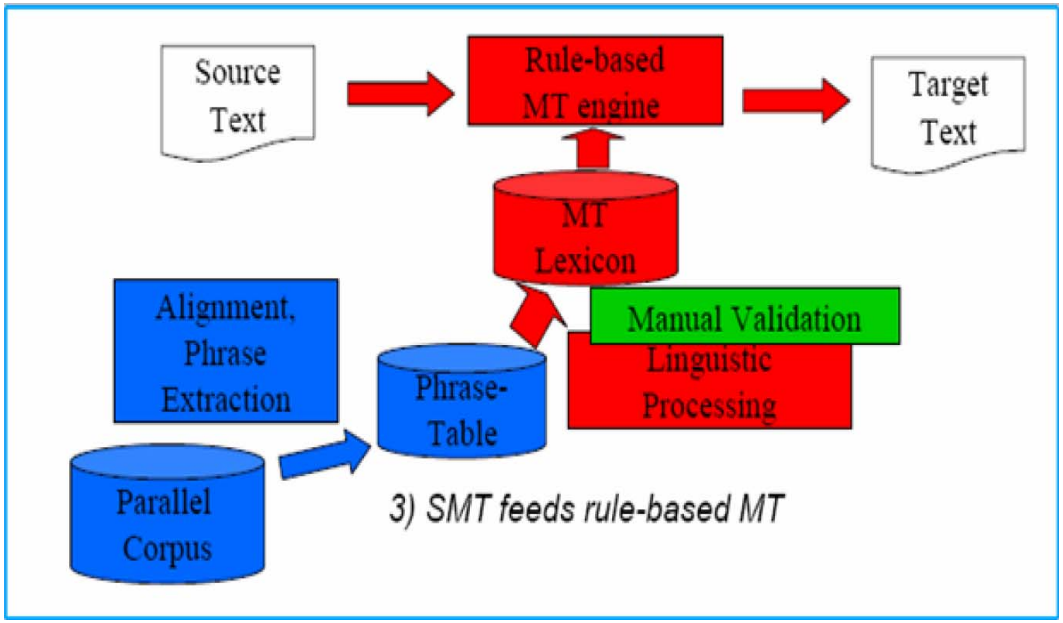
BUT:

- Works mainly for short utterances, where one of the candidates may be good enough (VerbMobil)
- Different candidates may have problems in different parts of the sentence, granularity of decisions too coarse

SMT feeds rule-based MT

Motivation:

- Adapting RBMT to new domains requires lots of new lexical entries that are difficult to write manually
- SMT techniques can help to partially automate this process



3) SMT feeds rule-based MT

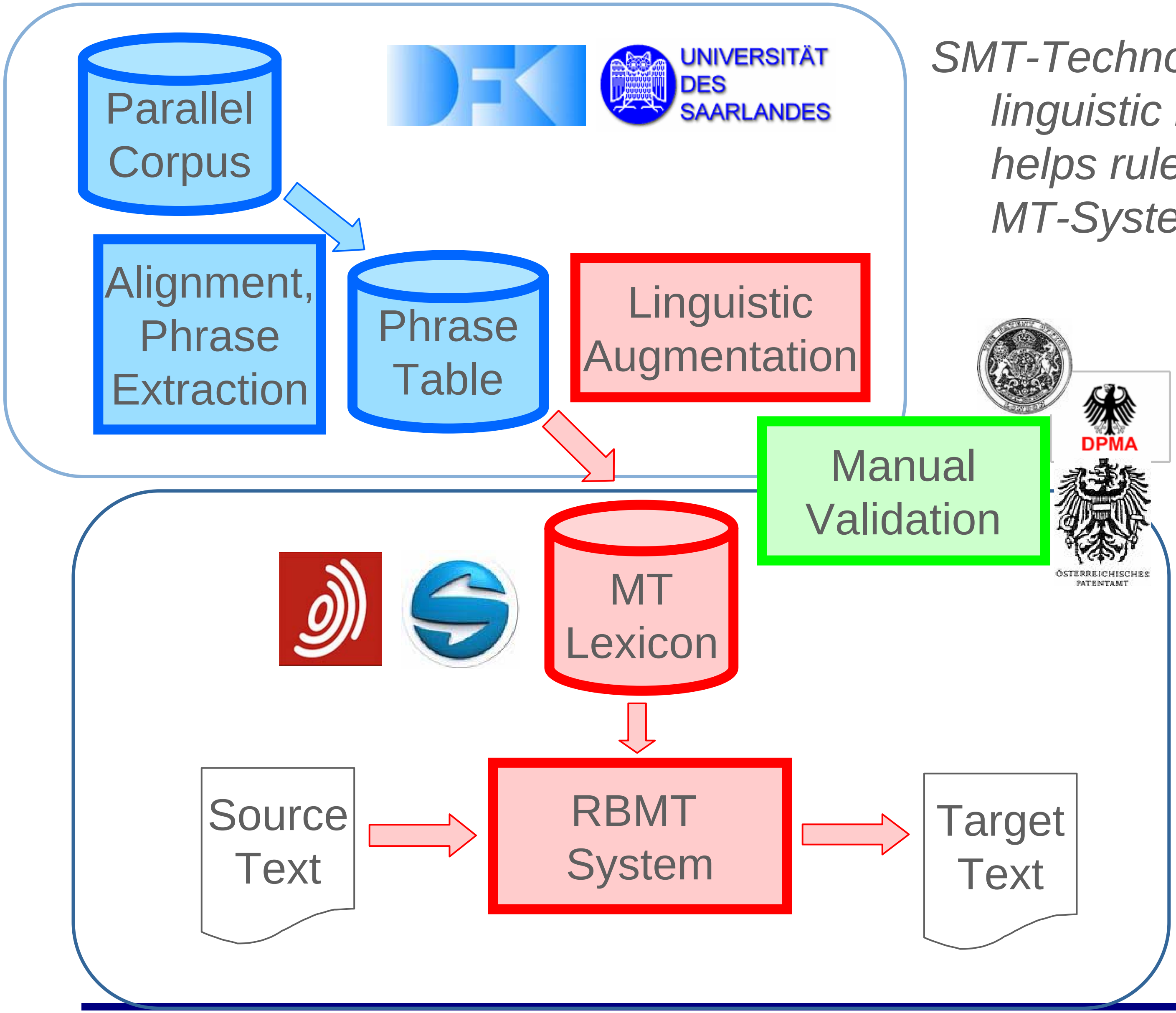
BUT:

- Not all required information can be learned from data
- Errors in examples/SMT alignment may creep in, but RBMT has no mechanism to discard implausible outcomes

➔ Some manual effort is required

Corpus-based Lexicon Extension for RBMT

SMT-Technology with linguistic knowledge helps rule-based MT-System



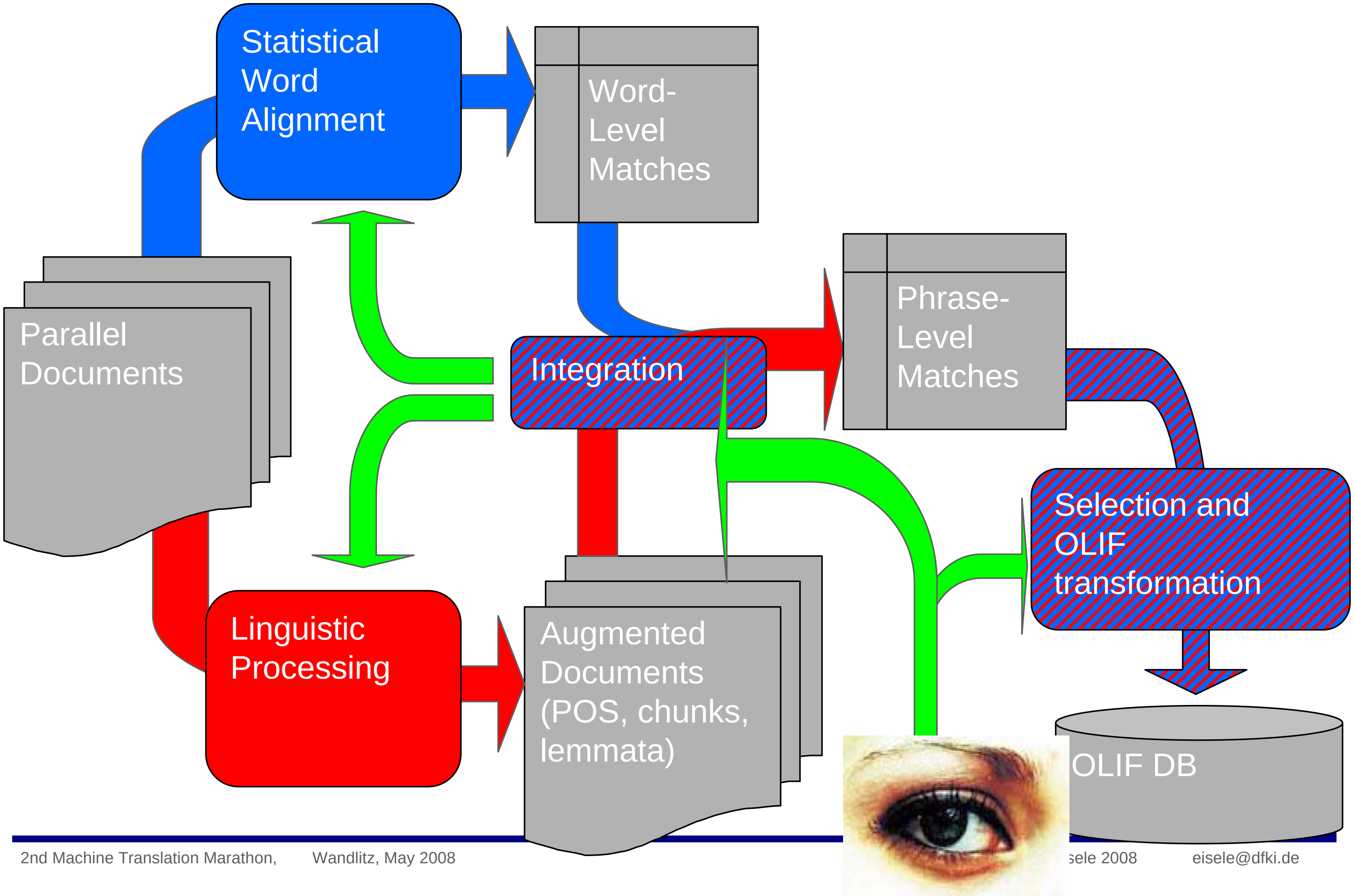
Language pairs

- ***DE ↔ EN***
- ***ES ↔ EN***
- ***FR ↔ EN***
- ***IT ↔ EN***

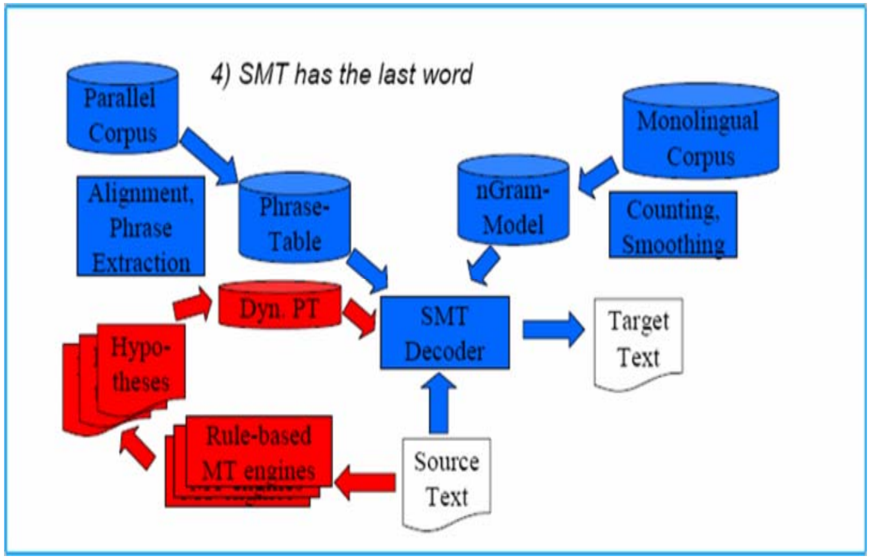
planned:

- ***EL ↔ EN***
- ***PT ↔ EN***
- ***NL ↔ EN***
- ***RO ↔ EN***
- ***FR ↔ DE***
- ***FR ↔ ES***

Terminology Extraction for MT: Architecture


Statistical Word Alignment
Word-Level Matches
Parallel Documents
Integration
Phrase-Level Matches
Selection and OLIF transformation
Linguistic Processing
Augmented Documents (POS, chunks, lemmata)
OLIF DB

RBMT feeds SMT



- Motivation: SMT can only know what is in the training data, RBMT systems often contain extensive lexical knowledge (e.g. Langenscheidt → T1 → Lucy)
- SMT decoder can be used to search for best combination of translation snippets from various sources

BUT:

Although architecture can fix lexical gaps, it but will not covercome problems with syntactically ill-formed candidates

RBMT feeds SMT

Current status:

- Preliminary version used in WMT07
- One completed diploma thesis, ongoing master‘s theses
- Generic implementation of alignment algorithm in a client-server setup, can be used for several other applications
- Promising results in WMT08:

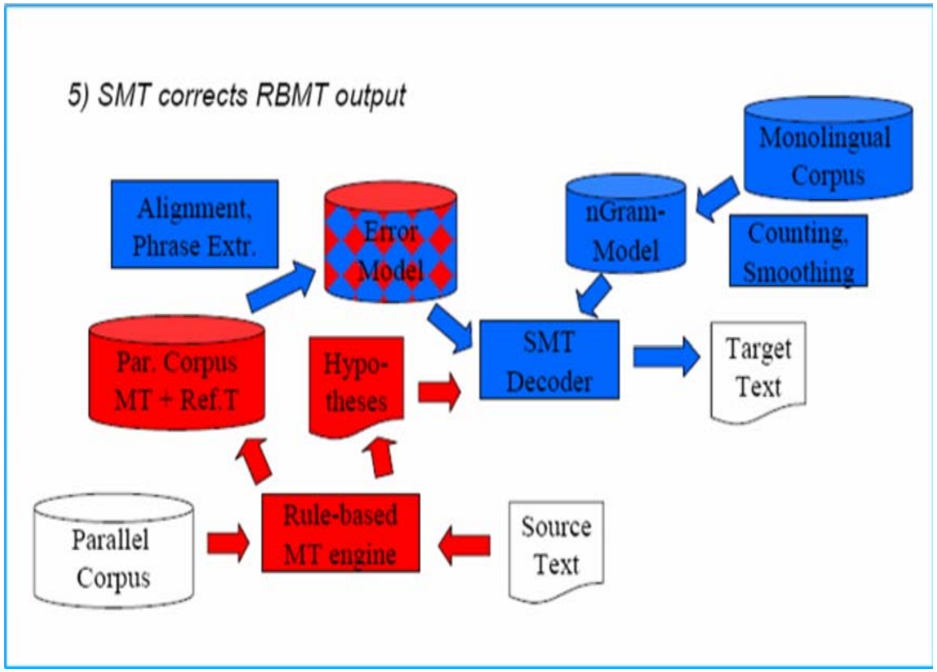
Ranks of USaar contribution relative to non-RBMT systems

	en-fr		en-de		en-es		fr-en		de-en		es-en	
	ep	nc	ep	nc	ep	nc	ep	nc	ep	nc	ep	nc
sentence ranking	2	3	5	1	3	3	3	5	2	1	6	1
yes/no	4	4	5	2	4	1	5	6	3	1	4	2
constituent ranking	4	2	4	2	2	2	5	7	1	1	1	3

RBMT feeds SMT: A working example

src	What did happen immediately after?
ref	Was geschah danach?
limsi	Was hat denn sofort nach?
liu	Was hat denn sofort nach?
uedin	Was geschah unmittelbar nach?
rbmt1	Was geschah sofort nachdem?
rbmt2	Was geschah nachher sofort?
rbmt3	Was geschah sofort danach?
rbmt4	Was geschah wirklich sofort danach?
rbmt5	Sofort nach was geschehen Sie?
rbmt6	Nachdem was sofort geschehen ist?
saar	Was geschah sofort danach?
sb-ct	Was geschah unmittelbar danach?

Statistical post-correction



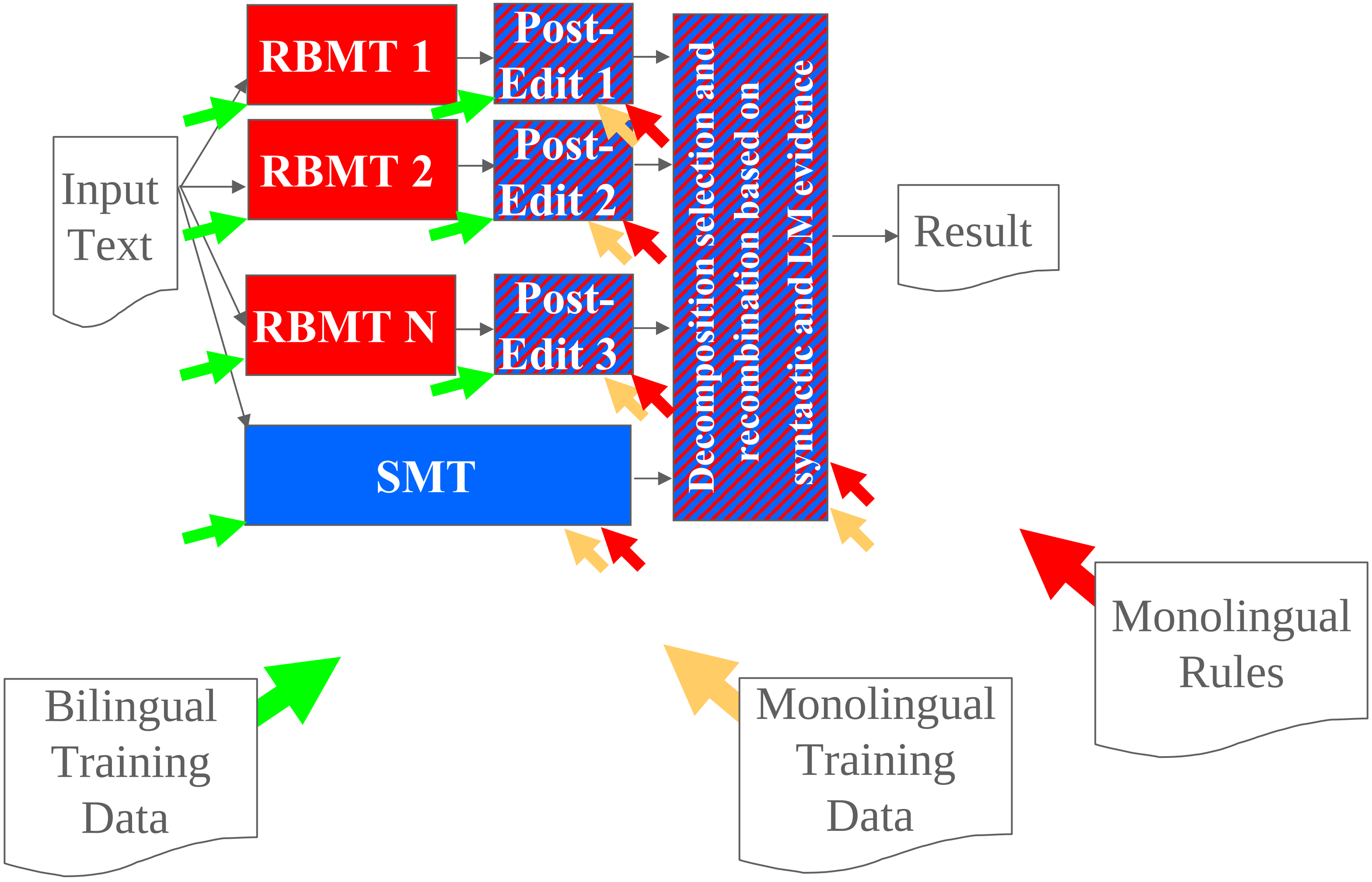
Motivation: Errors in RBMT can be systematic/regular, may be fixed automatically. Target language model helps to find most natural wording in context

This approach has show competitive results in recent work by UEdin, Systran, NRC, and LIMSI/Le Mans

BUT: Sometimes RBMT messes a sentence completely up, no hope to repair these cases via SMT. This can be alleviated by using multiple RBMT engines.

Competition vs. Integration

Ideas presented so far are independent, combinations are possible



Input Text

RBMT 1

RBMT 2

RBMT N

SMT

Post-Edit 1

Post-Edit 2

Post-Edit 3

Decomposition selection and recombination based on syntactic and LM evidence

Result

Bilingual Training Data

Monolingual Training Data

Monolingual Rules

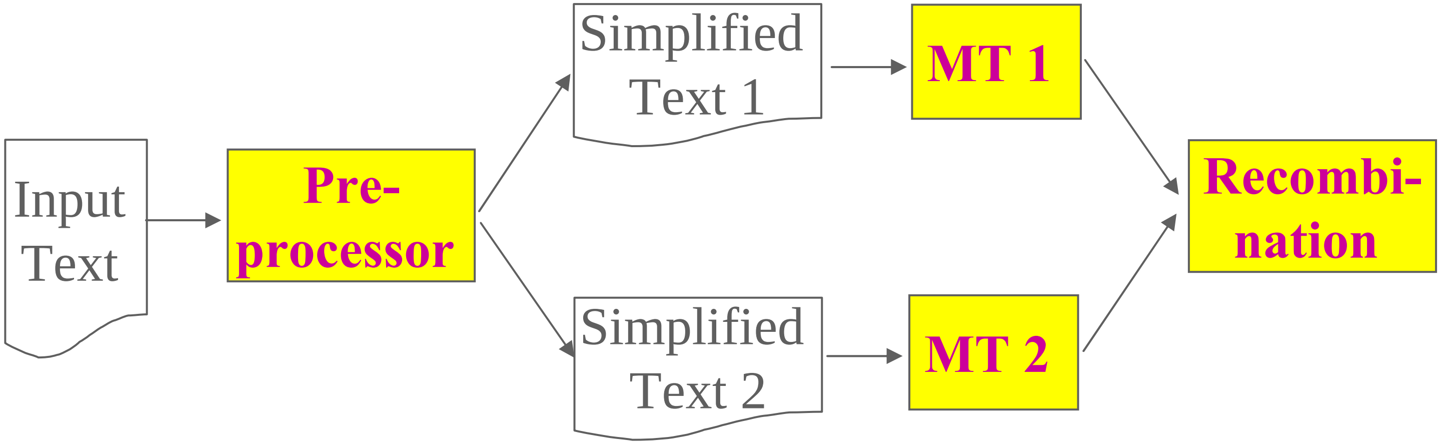
Pre-emptive division of labor between modules

The idea:

- So far, we send the input text unmodified through many MT systems, try to make sense of (partially erroneous) output
- Sometimes, a slight modification of the input can prevent errors from happening, e.g. by
  - replacing named entities unknown to the engine by place-holders
  - simplifying technical noun-phrases
- Statistics of error types can be used to find out specific weaknesses and best way to distribute work over engines

Pre-emptive division of labor

Schematic architecture

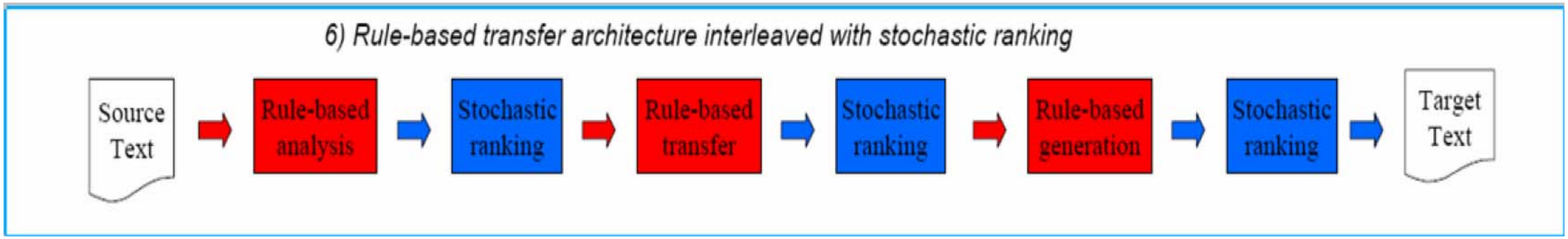


Actually already used in simplified form (e.g. for markup processing)

Open questions:

- Can we learn what to send through MT system from examples?
- What kind of pre-processing is adequate (should be robust ***and*** linguistically informed)

Transfer architecture with stochastic ranking



Motivation: Fine-grained combination of statistical and linguistic evidence on all levels requires a closely coupled implementation

BUT:

- Chain can only be as good as the weakest link
- Difficult to avoid mismatches between representations in hand-crafted grammars
- Many existing processing components are designed for deterministic processing; building up forests of alternative solutions may require redesign of algorithms

→ See talks by Petra Gieselmann, Stephan Oepen, and Micha Jellinghaus for work along these lines

Next Steps

- More careful analysis of WMT08 results, trying variants
- Systematic comparison between several hybrid approaches
  - RBMT→SMT vs. stochastic post-editing
  - Analyse impact of RBMT systems on quality of hybrid results
- Explore alternative approaches to system combination
- Error analysis, linguistic classification of problems
- Construction of stochastic models for important error types
- Identify ways to inspect intermediate representations and influence decisions within one RBMT system, e.g. Lucy

Conclusion

- Many different ideas of combining knowledge from RBMT and SMT systems have been presented, some of them have been successfully tried out
- Many of these approaches implement black-box integration, internals of RBMT do not have to be known
- These approaches seem to be independent, hence combinations are possible and should be evaluated
- Main drawback of system combination is the increase in overall complexity; effort should be seen as steps towards a unified architecture comprising all relevant knowledge sources