Transfer Rule Acquisition for MT

Michael Jellinghaus
Saarland University
micha@coli.uni-sb.de

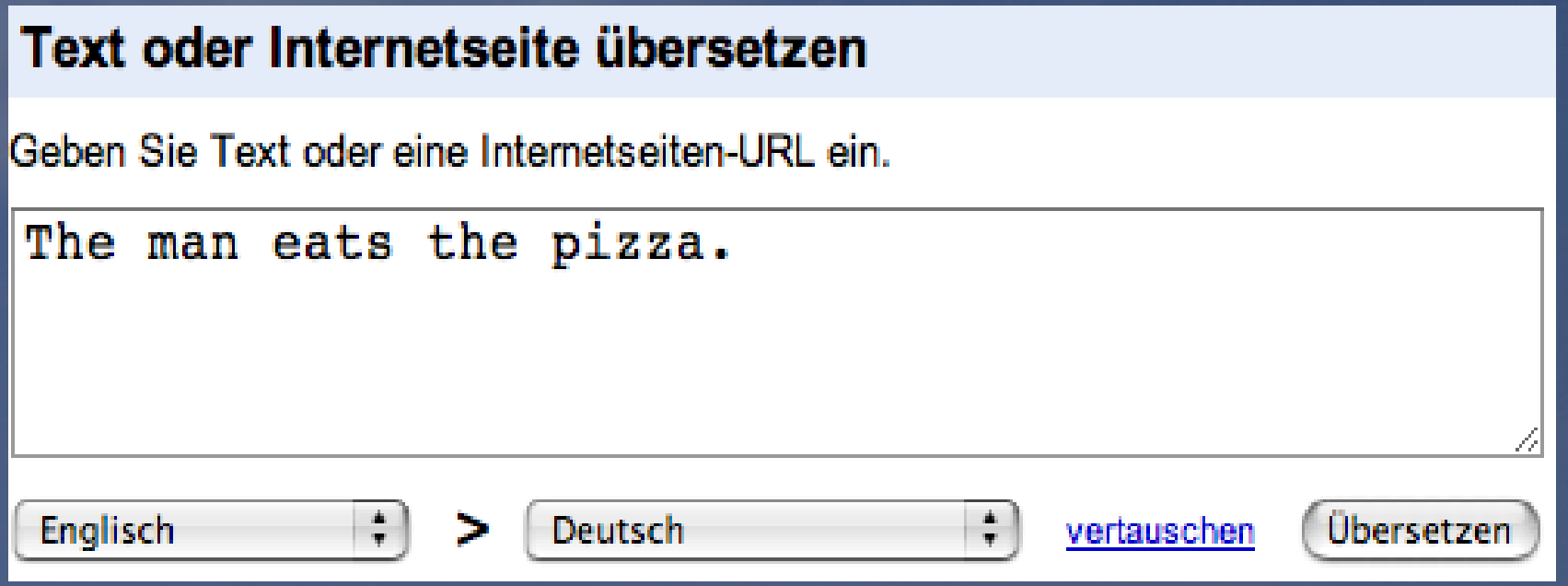
Overview

- Motivation
  - Characterization of Statistical and Transfer-Based Machine Translation
  - Hybrid System Idea
- Transfer Rule Acquisition
  - Parsing Details
  - Kinds of Rules
  - Rule Set Construction
- Issues / Discussion

Statistical Machine Translation

- Quick to develop
  - Translation model learned from parallel corpora
  - Target language model learned from monolingual corpora
- High coverage
  - Good lexical coverage (if items seen in training data)
    - *en: tax haven, de: Steueroase, fr: paradis fiscal*
    - *en: global warming, de: Erderwärmung, fr: réchauffement de la planète*
  - Robust: always delivers some output

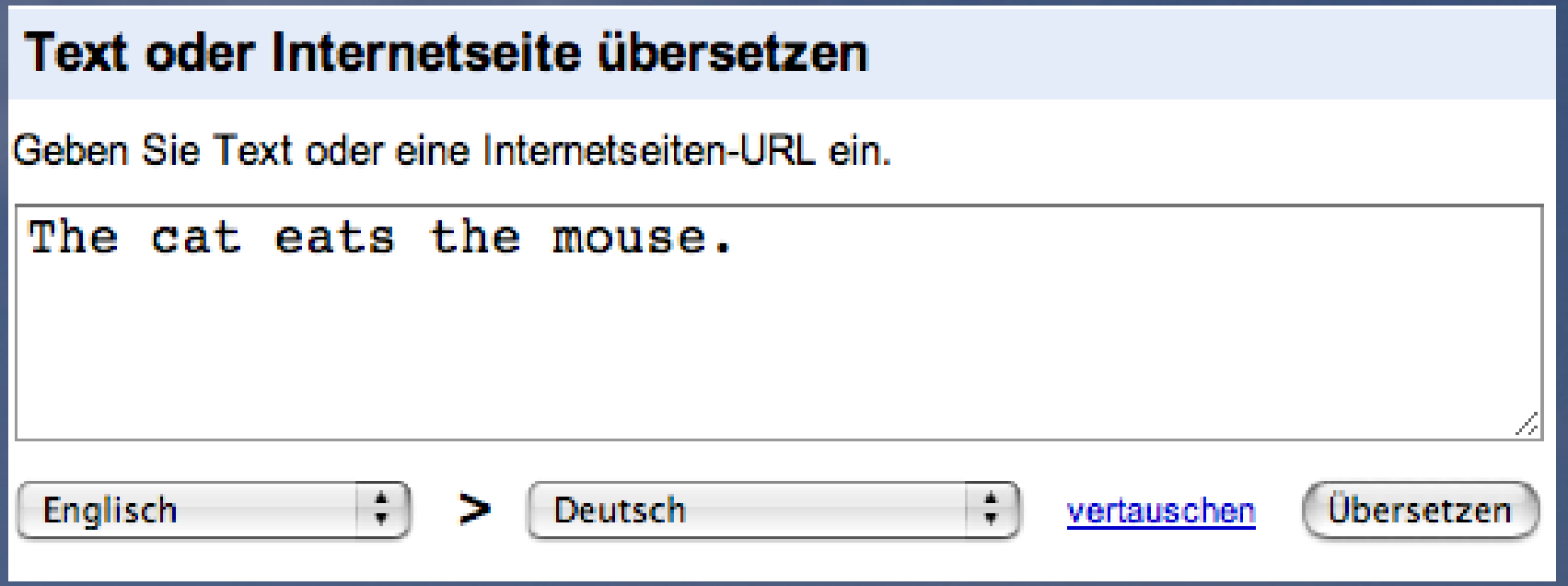
Statistical Machine Translation



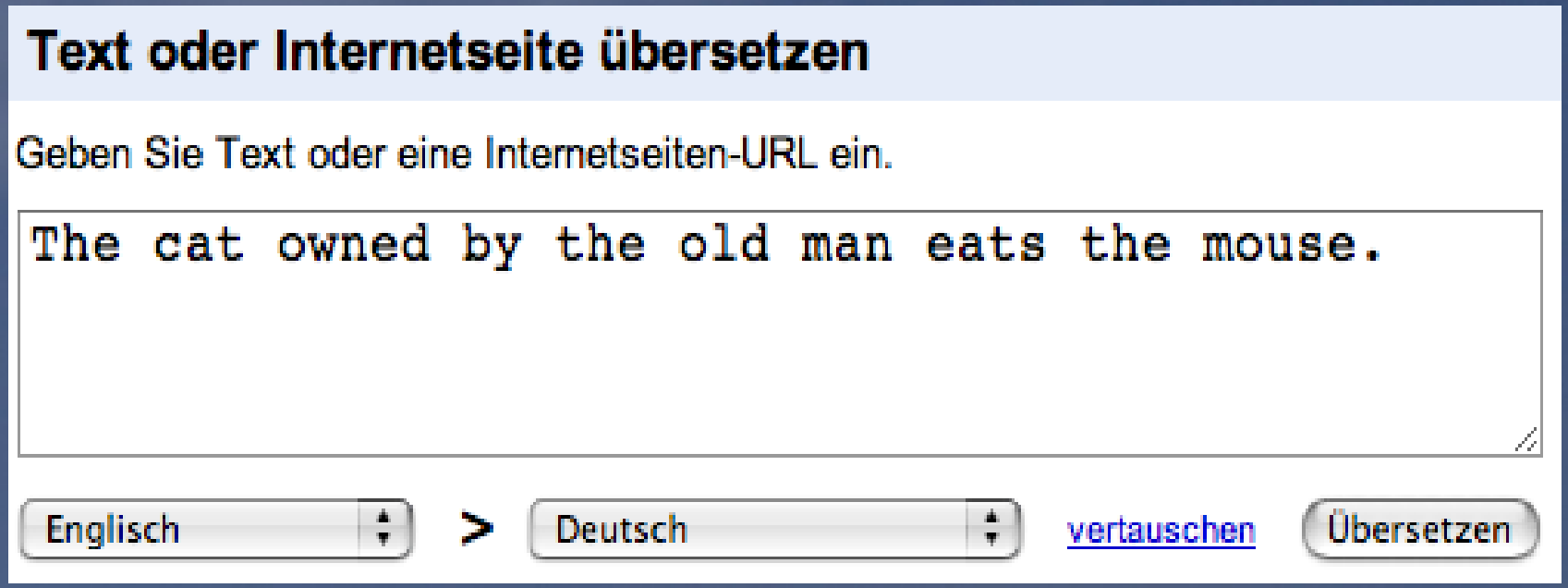
Übersetzung: Englisch » Deutsch

Der Mann isst die Pizza.

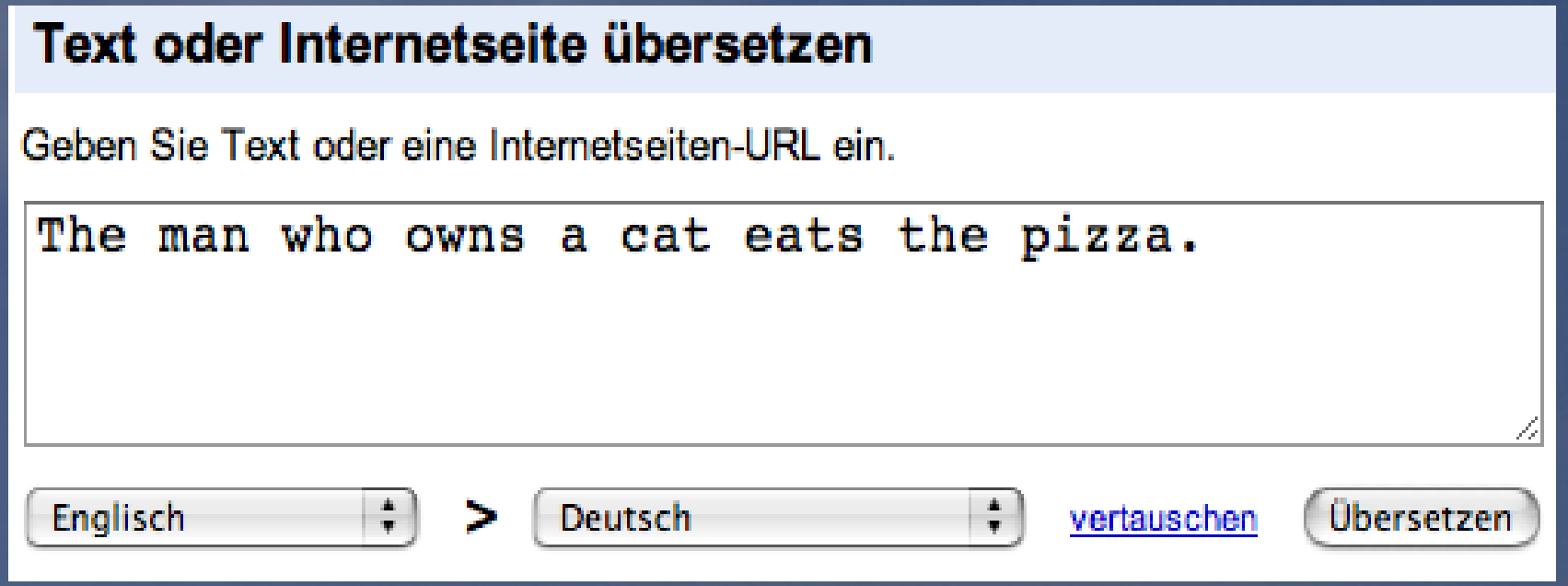
Statistical Machine Translation



Statistical Machine Translation



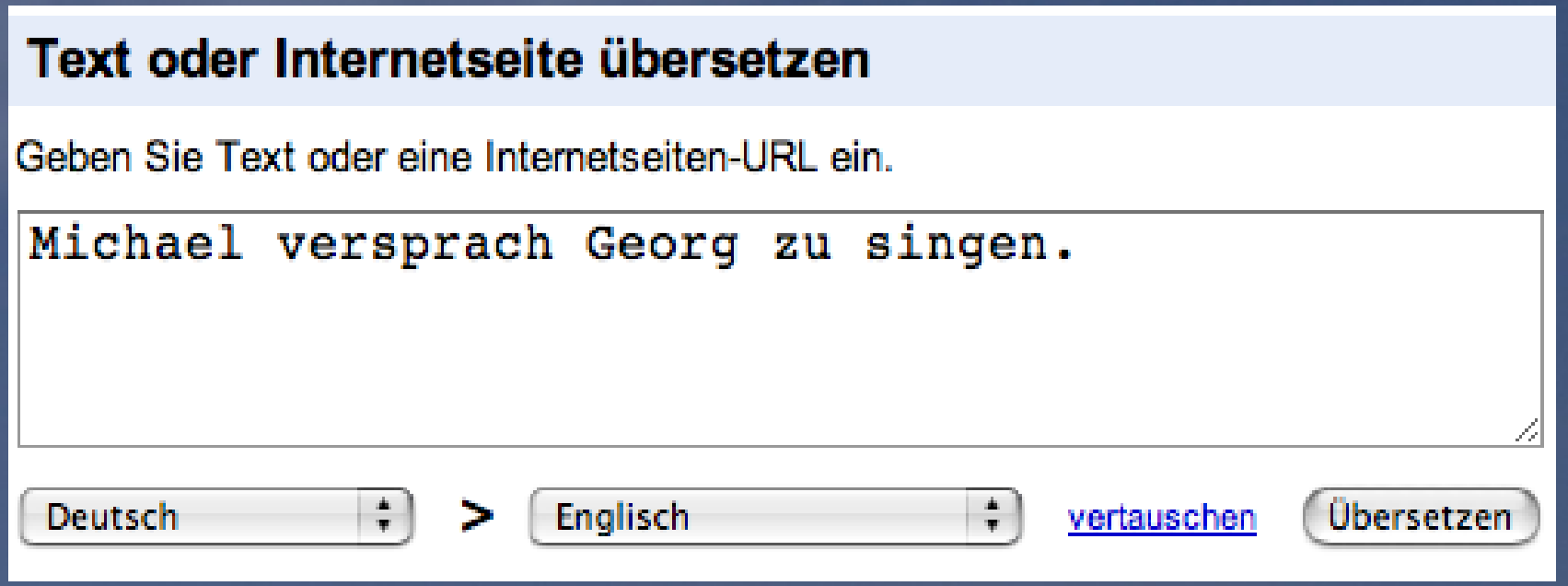
Statistical Machine Translation



Übersetzung: Englisch » Deutsch

Der Mann, besitzt eine Katze frisst die Pizza.

Statistical Machine Translation

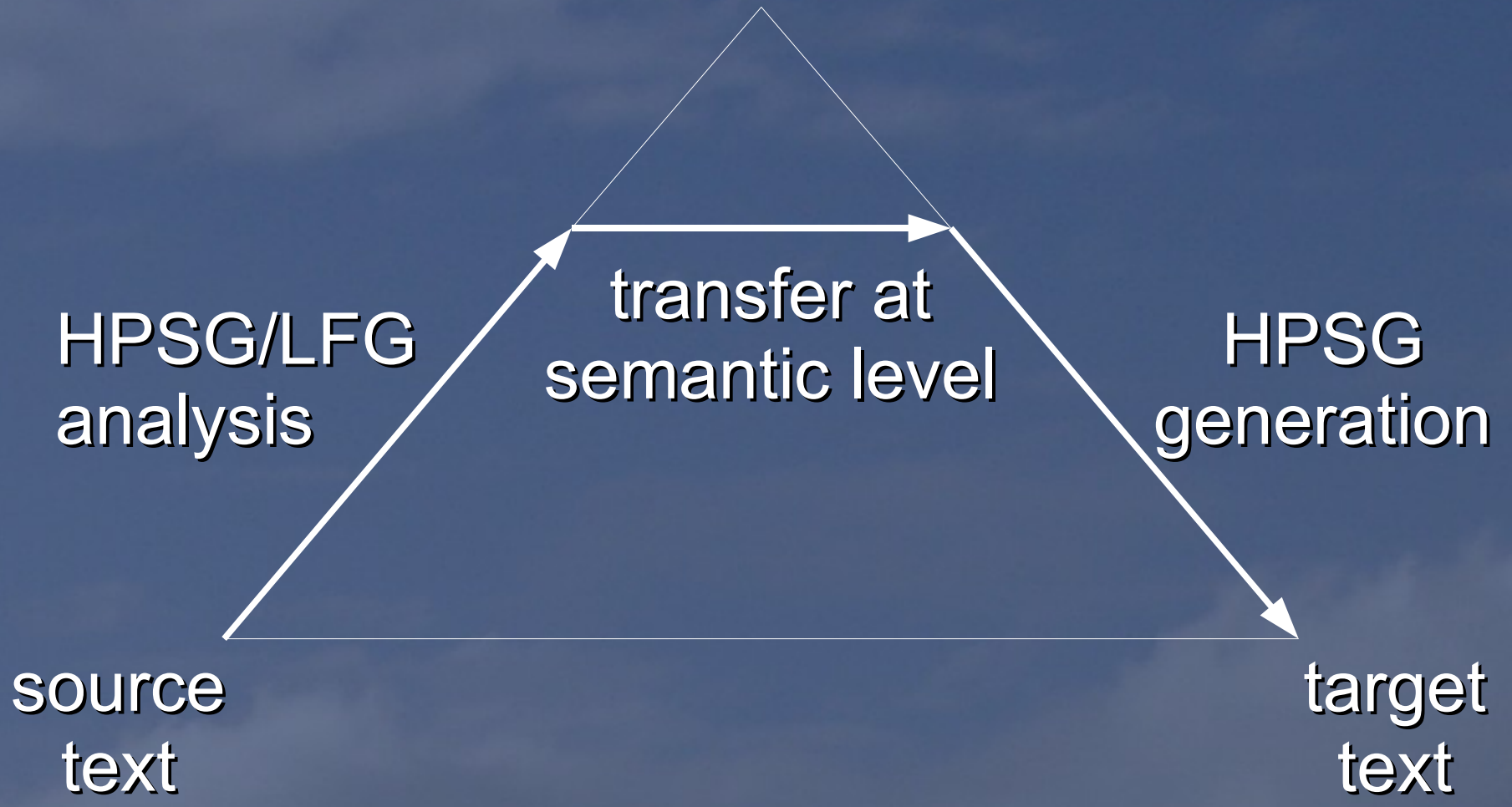


Advantages and Disadvantages

	SMT
development speed	+
grammaticality	-
lexical semantics	+
structural semantics	-
coverage	+

Deep Transfer-Based Machine Translation

MT using precision grammars: LOGON / DELPH-IN



Minimal Recursion Semantics

The cat eats the mouse. Analyze

results: all first | output: tree mrs | show: 5 results

[1 of 1 analysis; processing time: 0.01 seconds; 72 edges]

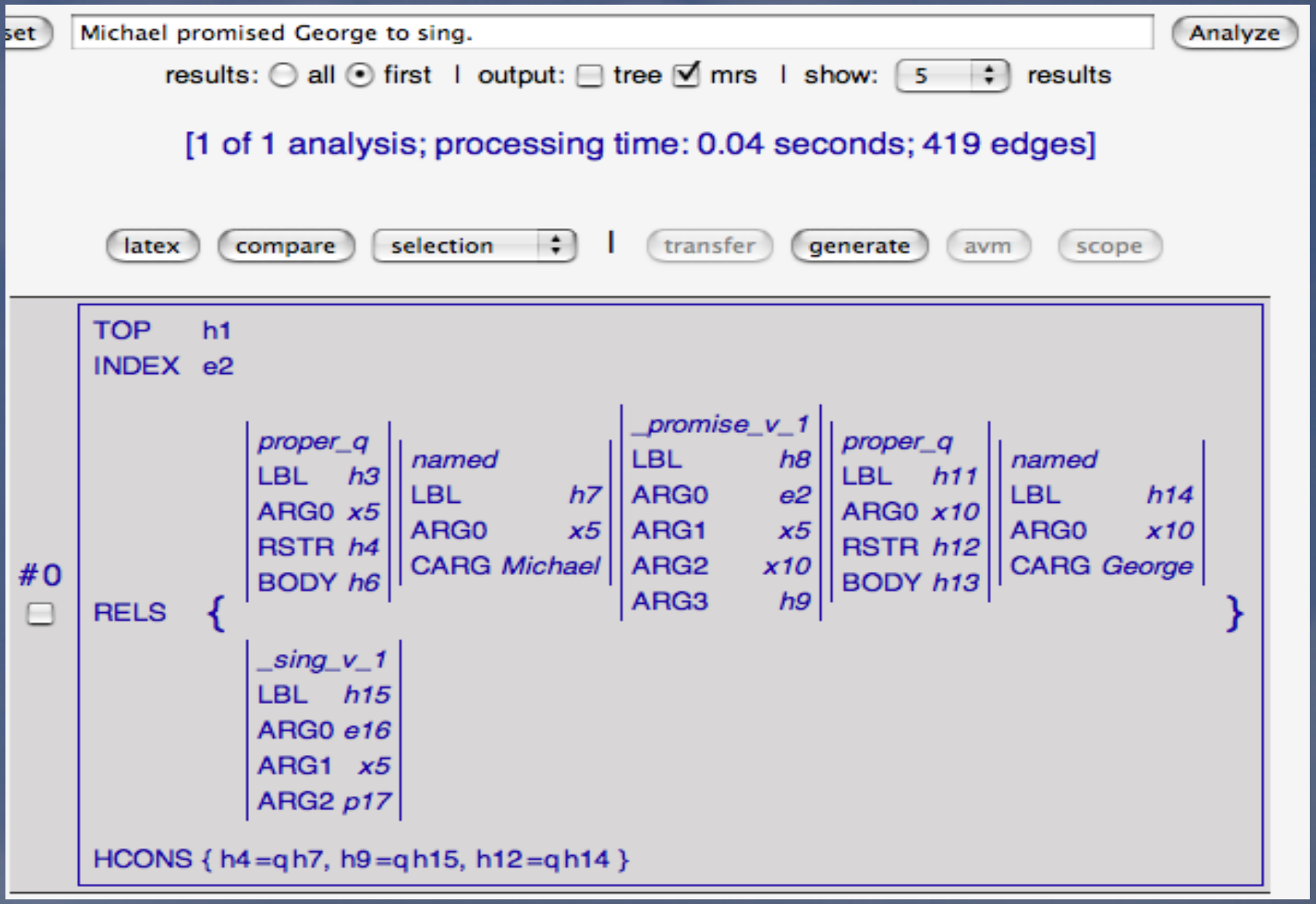
latex compare selection | transfer generate avm scope

#0

TOP	h1							
INDEX	e2							
RELS	{	*_the_q*		*_cat_n_1*		*_eat_v_1*	*_the_q*		*_mouse_n_1*	}
		LBL *h3*		LBL *h7*		LBL *h8*	LBL *h10*		LBL *h13*	
		ARG0 *x5*		ARG0 *x5*		ARG0 *e2*	ARG0 *x9*		ARG0 *x9*	
		RSTR *h6*				ARG1 *x5*	RSTR *h12*			
		BODY *h4*				ARG2 *x9*	BODY *h11*			

HCONS { h6 =q h7, h12 =q h13 }

Deep Transfer-Based Machine Translation



Deep Transfer-Based Machine Translation

- Advantages
  - Preserves meaning
  - Grammatical output
- Disadvantages
  - High development cost due to manual rule production
  - Output often not idiomatic
    - e.g. fr: *paradis fiscal* → en: *fiscal paradise*
  - Low coverage
    - e.g. *Steueroase* not in lexicon

Deep Transfer-Based Machine Translation

Advantages and Disadvantages

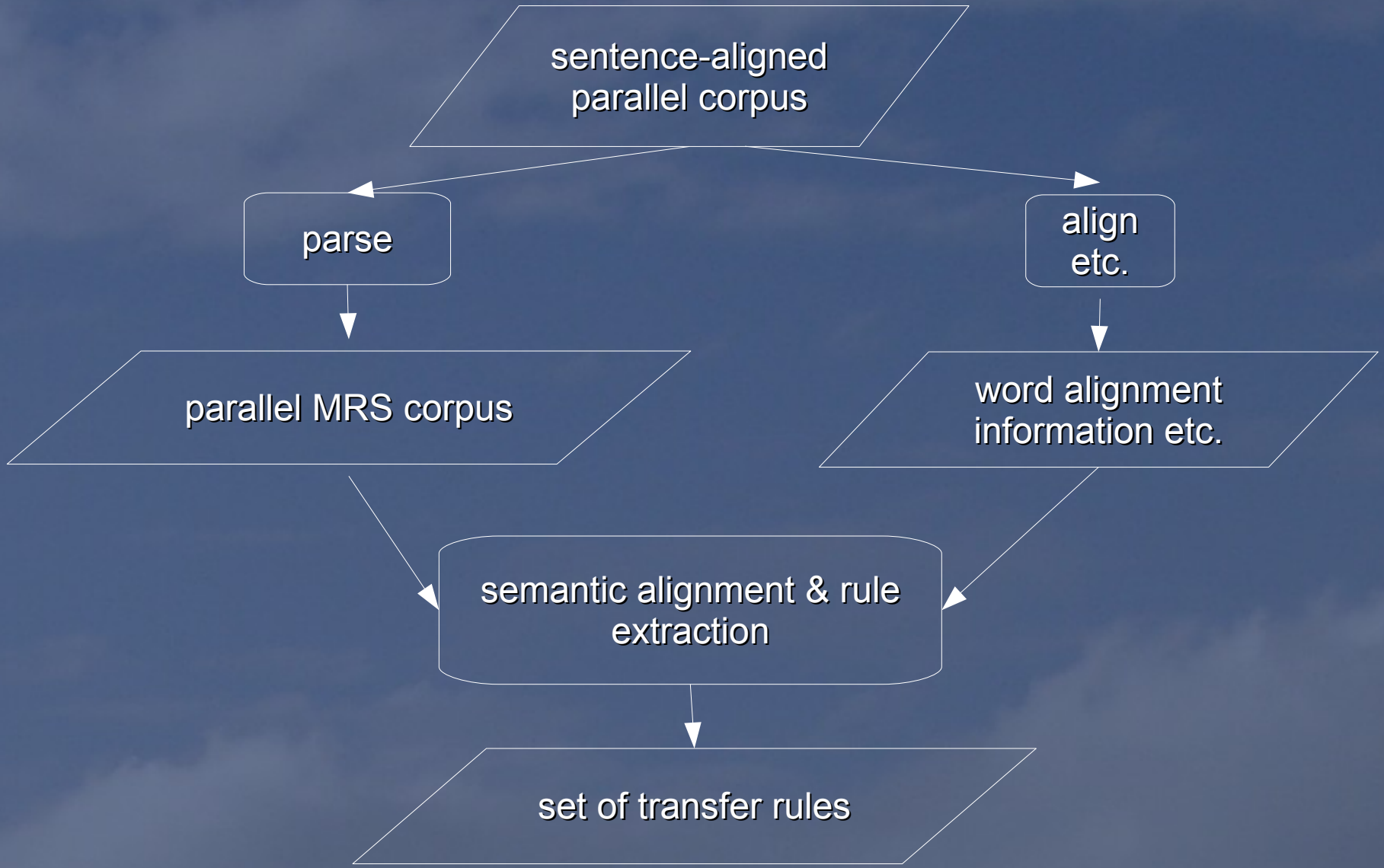
	SMT	DTBMT
development speed	+	-
grammaticality	-	+
lexical semantics	+	-
structural semantics	-	+
coverage	+	-

A Data-Driven Approach to Deep MT

- Idea: Combine complementary advantages by learning transfer rules from parallel corpora

	SMT	DTBMT	Hybrid
development speed	+	-	+
grammaticality	-	+	+
lexical semantics	+	-	+
structural semantics	-	+	+
coverage	+	-	+

Transfer Rule Acquisition Workflow



Parsing Details

- German: GG “Mar_2009”, decided to “freeze” as of 28 May 2009 (SVN revision 6575) due to trouble with new LKB version
- English: ERG “15-mar-07” (last one with message relations)
- Preprocessing: Tree tagger + preprocessor.fsr → YY format
- Parsing:

```
cheap -tok=yy -default-les -limit=100000 -packing
  -timeout=60 -mrs -tsdbdump ...
```

Simple “lexical” rules

```
katze_rule := monotonic_mtr &
[ INPUT [ RELS <
  [ PRED "_katze_n_rel", LBL #1, ARG0 #2 & [PERS #3, NUM #4] ] > ],
OUTPUT [ RELS <
  [ PRED "_cat_n_rel", LBL #1, ARG0 #2 & [PERS #3, NUM #4] ] > ] ].

named_rule := monotonic_mtr &
[ INPUT [ RELS < [ PRED "_named_n_rel", LBL #1, ARG0 #2, CARG #3 ] > ],
OUTPUT [ RELS < [ PRED named_rel, LBL #1, ARG0 #2, CARG #3 ] > ] ].

katalonien_rule := monotonic_mtr &
[ INPUT [ RELS < [ PRED "_named_n_rel", LBL #1, ARG0 #2, CARG "Katalonien" ] > ],
OUTPUT [ RELS < [ PRED named_rel, LBL #1, ARG0 #2, CARG "Catalonia" ] > ] ].
```

Simple “lexical” rules

```
singen_rule := monotonic_mtr &
[ INPUT [ RELS < [ PRED "_singen_v_rel",
  LBL #1, ARG0 #2 & [TENSE #3, MOOD #4], ARG1 #5 & [PERS #6, NUM #7] ] > ],
OUTPUT [ RELS < [ PRED "_sing_v_rel",
  LBL #1, ARG0 #2 & [TENSE #3, MOOD #4], ARG1 #5 & [PERS #6, NUM #7] ] > ] ].

gefallen_rule := monotonic_mtr &
[ INPUT [ RELS < [ PRED "_gefallen_v_rel", LBL #1, ARG0 #2, ARG1 #3, ARG2 #4] > ],
OUTPUT [ RELS < [ PRED "_like_v_rel", LBL #1, ARG0 #2, ARG1 #4, ARG2 #3] > ] ].

groß_rule := monotonic_mtr &
[ INPUT [ RELS < [ PRED "_groß_a_rel", LBL #1, ARG0 #2, ARG1 #3 ] > ],
OUTPUT [ RELS < [ PRED "_big_a_rel", LBL #1, ARG0 #2, ARG1 #3 ] > ] ].
```

Multi-word expressions / compounds

```
lispeln_rule := monotonic_mtr &
[ INPUT [ RELS < [ PRED "_lispeln_v_rel", LBL #1, ARG0 #2, ARG1 #3 ] > ],
OUTPUT [ RELS < [ PRED "_have_v_rel", LBL #1, ARG0 #2, ARG1 #3, ARG2 #4],
                [ PRED "_lisp_n_rel", LBL #5, ARG0 #4],
                [ PRED _a_q_rel, LBL h, ARG0 #4, RSTR #6, BODY h ] >,
         HCONS < qeq & [HARG #6, LARG #5] > ] ].

steueroase_rule := monotonic_mtr &
[ INPUT [ RELS < [ PRED "_steueroase_n_rel", LBL #1, ARG0 #2 ] > ],
OUTPUT [ RELS < [ PRED compound_rel, LBL #1, ARG0 e, ARG1 #2, ARG2 #3 ],
                [ PRED "_haven_n_rel", LBL #1, ARG0 #2 ],
                [ PRED "_tax_n_rel", LBL #4, ARG0 #3 ],
                [ PRED _udef_q_rel, LBL h, ARG0 #3, RSTR #5, BODY h ] >,
         HCONS < qeq & [HARG #5, LARG #4] > ] ].
```

Phrasal translations

```
mein_lieblingsbuch_rule := monotonic_mtr &
[ INPUT [ RELS < [ PRED "_lieblingsbuch_n_rel", LBL #1, ARG0 #2 ],
                 [ PRED "poss_rel", LBL #1, ARG0 e, ARG1 #2, ARG2 #3 ],
                 [ PRED "pronoun_q_rel", LBL h, ARG0 #3, RSTR #4, BODY h ],
                 [ PRED "pron_rel", LBL #5, ARG0 #3 ] >,
          HCONS < qeq & [HARG #4, LARG #5 ] > ],
OUTPUT [ RELS < [ PRED "_book_n_rel", LBL #1, ARG0 #2 ],
                  [ PRED "_like_v_rel", LBL #1, ARG0 #6, ARG1 #7, ARG2 #2 ],
                  [ PRED "_most_a_rel", LBL #1, ARG0 e, ARG1 #6 ],
                  [ PRED pronoun_q_rel, LBL #8, ARG0 #7, RSTR #9, BODY h],
                  [ PRED pron_rel, LBL #10, ARG0 #7 ] >,
           HCONS < qeq & [HARG #9, LARG #10 ] > ] ].
```

EP plus one or more of their arguments

```
the_cat_eats_rule := monotonic_mtr &
[ INPUT [ RELS < [ PRED "_eat_v_rel", LBL #1, ARG0 #2, ARG1 #3, ARG2 #4 ],
                 [ PRED the_q_rel, LBL #5, ARG0 #3, RSTR #6, BODY h ],
                 [ PRED "_cat_n_rel", LBL #7, ARG0 #3 ] >,
          HCONS < qeq & [HARG #6, LARG #7 ] > ],
OUTPUT [ RELS < [ PRED "_fressen_v_rel", LBL #1, ARG0 #2, ARG1 #3, ARG2 #4 ],
                  [ PRED "def_q_rel", LBL #5, ARG0 #3, RSTR #6, BODY h ],
                  [ PRED "_Katze_n_rel", LBL #7, ARG0 #3 ] >,
           HCONS < qeq & [HARG #6, LARG #7 ] > ] ].
```

Rules for other co-occurrence effects

```
großer_mann_rule := monotonic_mtr &
[ INPUT [ RELS < [ PRED "_groß_a_rel", LBL #1, ARG0 #2, ARG1 #3 ],
                 [ PRED "_mann_n_rel", LBL #1, ARG0 #3 ] > ],
OUTPUT [ RELS < [ PRED "_tall_a_rel", LBL #1, ARG0 #2, ARG1 #3 ],
                [ PRED "_man_n_rel", LBL #1, ARG0 #3 ] > ] ].
```

Construction of Transfer Rule Set

Quality control:

For a given sentence pair, the extracted rules are only retained if the semantic structures on both sides could be aligned completely.

Construction of Transfer Rule Set

- All rules are added to one big set of rules
- Their order in the rule set determines the order in which their application is tried
- When applied, rules consume their input which is then no longer available as input to other rules

→ Sort rules by number of input EPs
("specific rules first" strategy to capture co-occurrence information)

Construction of Transfer Rule Set

- Rules can be optional
- For performance reasons, not all should be

→ For a given input side, the last rule is always mandatory

→ Extraction frequency as a second sorting criterion (in order to eliminate noise)

Examples of noise:

- *Wer...* → *what group...* (loose translation)
- *Das Kind jagt die Katze* (ambiguity)
- Other errors at the various levels (parsing, alignment, ...)

Issues

- Cyclic structures in some MRSs
- Punctuation in CARG values
  - *Wofür steht TMB? → What does TMB? stand for?*
- Maximum size of rule set?
- Parse selection (for rule extraction step)
- Generalization strategies?

Thank you!

Questions or comments?