The German to English METIS-II MT System

Michael Carl,

IAI

Overview:

- METIS-II: General Introduction:
  - Project description
  - Architecture of German to English
- Detailed Description:
  - Source language analysis
  - Dictionary matching and lookup
  - Target language adjustment
  - Translation ranking and selection
  - Target language token generation.

METIS-II: Time Table

- EU Project within IST-STREP
- Started October 2004
- Duration: 3 Jahre
- Until June 2006:
  Exploration of various methods
- Since June 2006:
  refinement/integration of modules
  interface for user adaptation

Participants

- ILSP, Athen: Greek —-> English
- CCL, Leuven: Dutch —-> English
- IAI, Saarbrücken: German —-> English
- UPF, Barcelona: Spanish —-> English

METIS-II: Goals

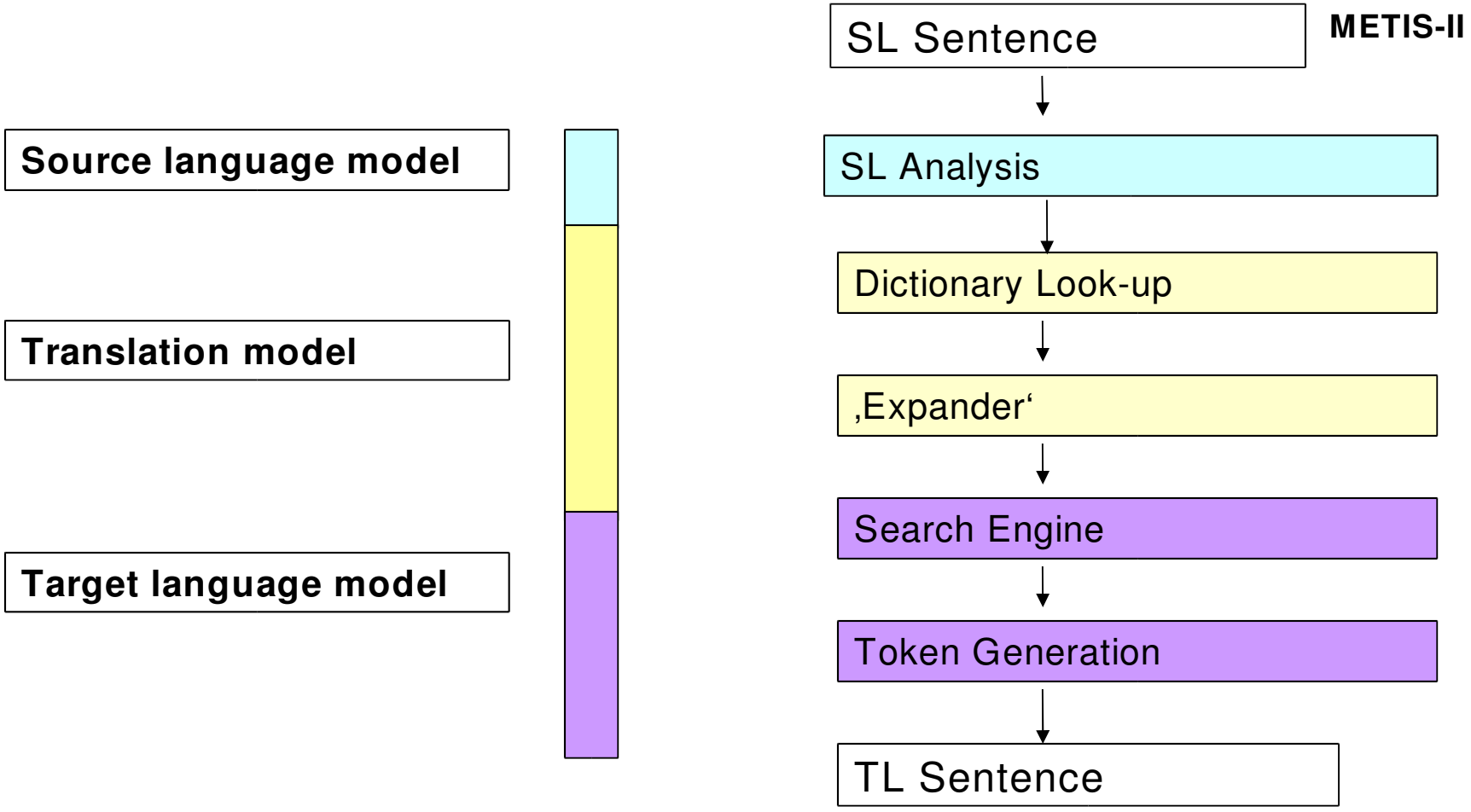
METIS-II

- METIS-II is the continuation of METIS-I:
  - exploit entities below sentence border
- METIS-II resources:
  - use only ‘basic' tools and resources
- METIS-II does not need a particular format:
  - enable different tag-set for SL and TL
  - plug in different taggers
- METIS-II can be used for several languages

Required Resources

- Bilingual Dictionary:
  - German to English
- Basic 'linguistic' tools:
  - SL and TL tagger, chunker
- Monolingual TL Corpus:
  - BNC (10^8 words, 10^6 sentences)
- Parallel Corpora (SMT/EBMT) **not** required
  - avoid data-acquisition bottleneck

Overview of the System



SL Model: German Analysis

- MPRO:
  - lemmatization
  - morphological analyser
- KURD / FRED (shallow syntax analysis):
  - grammar is basis for:
    - Duden Korrektor (German grammar checker)
    - CLAT (Controlled language technology)
    - text indexation
    - pattern-based formalism to detect and mark phrases, clauses, topological fields

German Grammar

METIS-II

- Recognised Constituents (flat representation):
  - NPs, PPs, Verbal groups
  - clauses
  - topological fields
  - does not detect/mark relation between constituents
- Method:
  - originates in requirements for grammar correction
  - iterative process:
    - mark 'secure' patterns
    - disambiguate the pattern

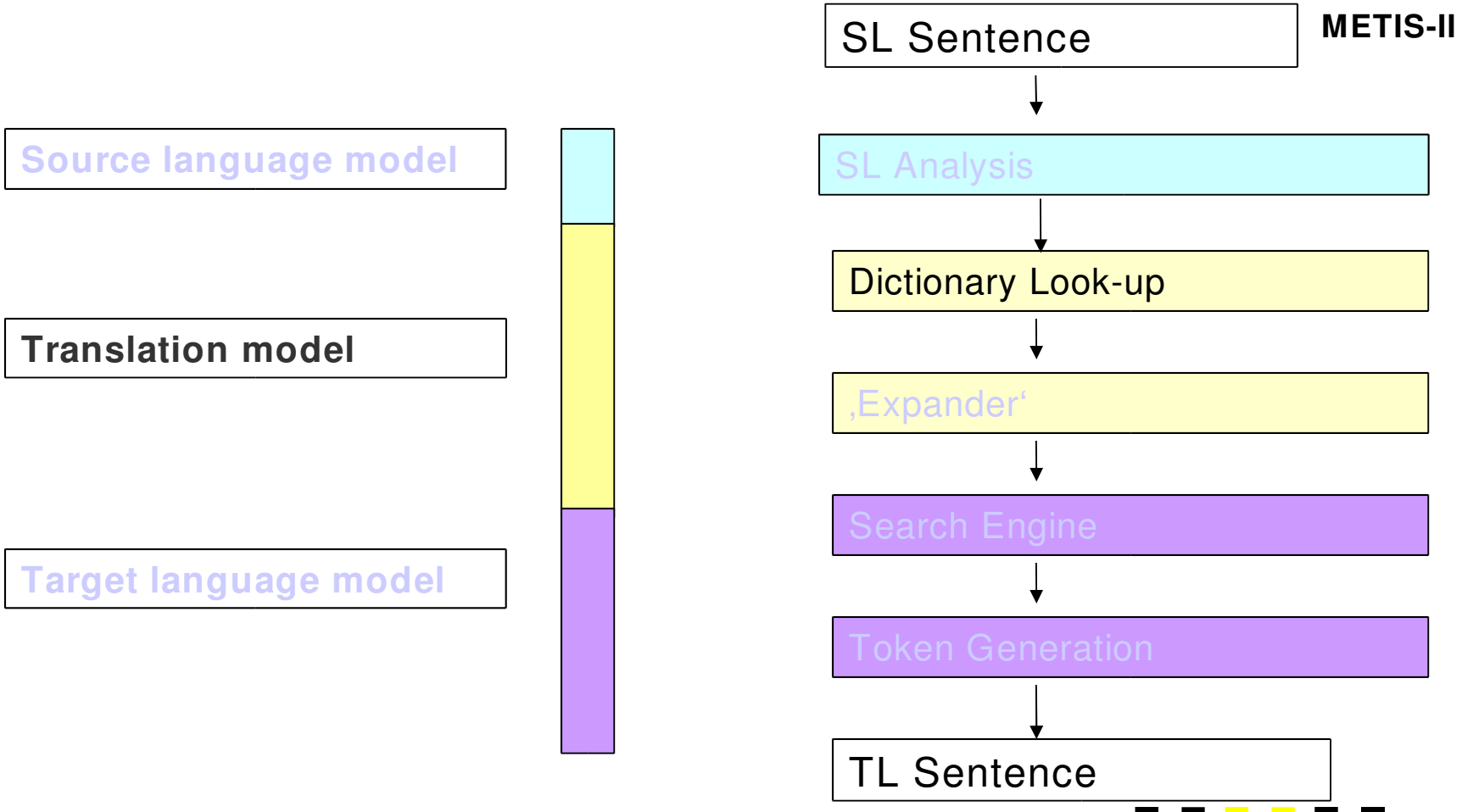
Input/Output of German Analyser

- Das Haus wurde von Hans gekauft
  The house was from Hans bought

Lemma	wnr	PoS	phrase	clause/field
{lu=das,	wnr=1,	c=w,sc=art,	phr=np;subjF,	cl=hs;vf} ,
{lu=haus,	wnr=2,	c=noun,	phr=np;subj,	cl=hs;vf},
{lu=werden,	wnr=3,	c=verb,vt=fiv,	phr=vg fiv,	cl=hs;lk},
{lu=von,	wnr=4,	c=w,sc=p,	phr=np;nosubjF,	cl=hs;mf},
{lu=Hans,	wnr=5,	c=noun,	phr=np;nosubj,	cl=hs;mf},
{lu=kaufen,	wnr=6,	c=verb,vt=ptc2,	phr=vg ptc,	cl=hs;rk}

.

Translation Model: Dictionary Look-up



German-to-English Dictionary

METIS-II

- > 600.000 Entries
- Independent tag sets in SL and TL
- Single- and multi word units, phrase translations
- Represented as flat trees:
  - leaves contain lexical information
  - mother node contains meta information

Goals of Dictionary Look-up

- Discontinuous Entries:
  - separable prefix — lehnt ... ab <--> reject
  - reflexive verbes — sich ... beeilen <--> hurry up
  - support verbes — in Gefahr bringen <--> endanger
  - idioms — vom Mund ablesen <--> lip-read
- Lexical Overgeneration:
  - lex.-sem. ambiguities — Bank <--> bank;bench
  - main/aux.verb — werden <--> will;be;become
  - negation — nicht <--> do not;not
  - magniers/intensifers — stark <--> strong;good;heavy ...
  - prepositions — auf <--> on;in;up;onto ...

Types of Discontinuous Verbal Realisations

METIS-II

- Dictionary entry:
  - **Anweisung ausführen** <--> *execute statement*
- Realisation in a subordinate clause (en bloc):
  - dass er sofort die **Anweisungen ausführt** …
    *that he immediately the **statements executes** …*
- Realisation in a main clause (left Klammer & Mittelfeld) :
  - Er **führt** die **Anweisung** sofort **aus** …
    *He **executes** the **statement** immediately **VPREF ...***
- Realisation in a modal main clause (Mittelfeld & right Klammer):
  - Er will die **Anweisung** sofort **ausführen.**
    *He will the **statement** immediately **execute.***

Dictionary Maintainance and Look-up

METIS-II

- Structure and Maintainance of the dictionary:
  - lemmatisation and morphological analysis of entries
  - consistency of entries
  - generation of variants
  - indexation of morphemes
- Dictionary look-up:
  - retrieve entries and filter 'best' matches
    - lexical similarity
    - contextual consolidation

Structure of Dictionary Entry

{de=einsperren,mde={c=verb}, en=lock_<so.>_away,men={c=verb}}.
{de=ausführen, mde={c=verb}, en=execute,men={c=verb}}.

- Structure of dictionary entries:
  - represented as flat trees
  - contain lexical information and meta information
  - leaves follow canonical representation
- Problem:
  - inflexion, derivation, variation
    ==> lemmatization, analyse morphological structure

Canonical Forms of Dictionary Entries (German side)

- c=verb: last word of entry is infinite verb
  e.g. “tanzen gehen” *(dancing go)*
- c=noun: last word of entry is noun/sing/nominative
  e.g. “dritte Welt“ *(third world)*
- c=adj: last word of entry is adjective:
  e.g. “hell grün“ *(light green)*
- c=p: last word of entry is preposition
  e.g. “in Bezug auf“ *(with respect to)*

Morphological Analysis of German Entries

{de=ausführen, mde={c=verb},en=execute,men={c=verb}}.

“ausführen” has morphological sturcture: “ls=aus_$führen“

and several morphological analyses:

{lu=ausführen,c=noun,ehead={nb=sg,case=acc;dat;nom,g=n}};
{lu=ausführen,c=verb,vtyp=fiv,nb=plu,per=1;3,tns=pres};
{lu=ausführen,c=verb,vtyp=inf}.

Disambiguated entry (according to canonical form):

{c=verb}@{lu=ausführen,ls=aus_$führen,c=verb}.

Lexical Variation

METIS-II

- Abfertigung des Gepäcks --> Gepäckabfertigung
  check-in of the luggage --> lugguage check-in

- Anzahl der Mitarbeiter --> Mitarbeiteranzahl
  number of worker --> *worker number*
  {c=noun}@{c=noun,ls=anzahl},{c=art,ls=art},{c=noun,ls=mit_$arbeiten}.
  -->
  {c=noun}@{c=noun,ls=mit_$arbeiten#anzahl}.

- ausführen ☮ führen ... aus
  {c=verb,type=ns}@{c=verb,ls=aus_$führen}.
  -->
  {c=verb,type=hs}@{c=verb,ls=führen},{c=vpref,ls=aus}.

Dictionary-lookup: Retrieval and filtering

METIS-II

- Retrieve entries which share morphological structure
- Filter best candidates:
  - consolidate word order
    - dictionary entry and match in same word-order
  - compute lexical delta
    - find most 'similar' word form
  - contextual consolidation
    - check 'internal' and external context of match

Some Surface Realisations of „aus_$führen“

Word	Lemma	PoS	Derivation	Feature Info
Ausführbarkeit	ausführbarkeit	noun	~bar~heit	nb=sg
Ausführer	ausführer	noun	~er	nb=sg
Ausführung	ausführung	noun	~ung	nb=sg
ausführen	ausführen	verb	---	per=1;3, tns=pres
ausführbar	ausführbar	adv	~bar	deg=base
ausführbarer	ausführbar	adv;adj	~bar	deg=comp
ausgeführten	ausgeführt	adj	ptc2	deg=base
ausgeführteren	ausgeführt	adj	ptc2	deg=comp
ausgeführt	ausführen	adj;verb	ptc2	
Ausführender	ausführend	adj	ptc1	deg=base
ausführend	ausführend	adv	ptc1	

Lexical delta for Morph. Structure: aus_$führen

METIS-II

- Dictionary entries:
  - ausführen <--> export (verb)
    {lu=ausführen,c=verb,nb=plu,per=1;3,tns=pres}
  - ausgeführt <--> executed (participle)
    {lu=ausgeführt,c=adj,ptc=2,deg=base}
- Inflected German forms in sentence:
  - ausführst (inflected verb)
    {lu=ausführen,c=verb,nb=sg,per=2,tns=pres}
    --> match: ausführen <--> export
  - ausführende (present participle)
    {lu=ausführend,c=adj,ptc=1,deg=base}
    --> match: ausgeführt <--> executed

Contextual Consolidation of 'verbal' Entries (1)

METIS-II

- ***Anweisung ausführen*** *<--> execute statement*

  main clause:

  – If ((verbal part of entry is left Klammer) and
    (**nominal part** of entry is in Mittelfeld)) then
    consolidate match
    end

  – Er **führt** die **Anweisung** sofort **aus**
    *He **executes** the **statement** immediately **VPREF** ...*

Contextual Consolidation of 'verbal' Entries (2)

METIS-II

- ***Anweisung*** ausführen <--> *execute statement*

  modal main clause:

  – If ((verbal part of entry is right Klammer) and
    (**nominal part** of entry is in Mittelfeld)) then
    consolidate match
    end

  – Er will die **Anweisung** sofort **ausführen**.
    *He will the **statement** immediately execute.*

Contextual Consolidation
Nominal Entries

METIS-II

- *Abbau der Ozonschicht <--> depletion of ozone*

  within a noun phrase (np):

  – *Only additional adjectives may modify the entry:*

  Abbau der arktischen Ozonschicht
  depletion of arctic ozone

Output of Dictionary Look-up

METIS-II

{lu=das,wnrr=1,c=w,sc=art, ... }
@{c=art,n=146471}@{lu=the,c=AT0}. .
,{lu=Haus,wnrr=2,c=noun, ...}
@{c=noun,n=268244}@{lu=company,c=NN1}.
, {c=noun,n=268246}@{lu=home,c=NN1}.
, {c=noun,n=268247}@{lu=house,c=NN1}.
, {c=noun,n=268249}@{lu=site,c=NN1}. .
,{lu=werden,wnrr=3,c=verb,vtyp=fiv, ...}
@{c=verb,n=604071}@{lu=be,c=VBD} .
, {c=verb,n=604076}@{lu=will,c=VM0} . .
...

Discontinuous Match

METIS-II

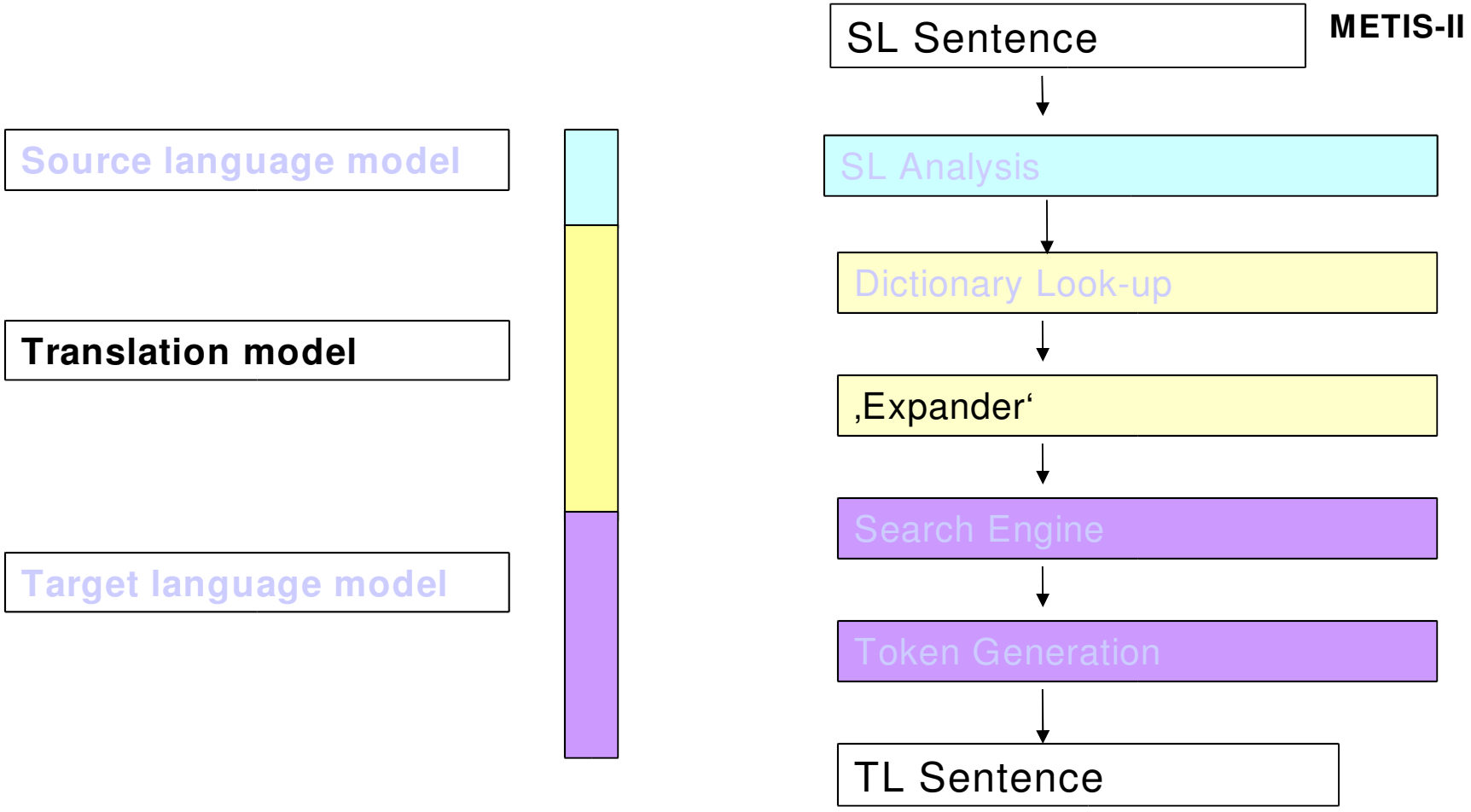
Das **geht**, solange es Frauen gibt, nie **vor die Hunde**.

vor die Hunde gehen <---> go to the dogs | be buggered

,{lu=gehen|...|vor|der|hund,wnrr=2;10;11;12,c=verb,markcl=hs}
@{c=verb,n=13}@{lu=go,c=VVB;VVD;VVI;VVN;VVZ}
, {lu=to,c=TO0;PRP}
, {lu=the,c=AT0}
, {lu=dog,c=NN2;NN1} .
, {c=verb,n=14}@{lu=be,c=VBB;VBD;VBI;VBN;VBZ}
, {lu=bugger,c=VVN;VVD} . .

...

Translation Model: Expander



Expander

METIS-II

- Rule-based device to adjust word order
- insert/delete/permute/modify translation hypotheses in AND/OR graph:
- insert article — Hans ist Lehrer --> Hans is **a** teacher
- verbal group — Das Haus **wurde** von Hans **gekauft**
  --> The house **was bought** by Hans.
- add hypotheses — Die Milch trinkt die Katze.

--> (The cat drinks the milk. | The milk drinks the cat.)
Peters Auto --> (Peters' car | the car of Peter)

Example of an Expander Rule

Hans **hat** das Haus **gekauft**. --> Hans **hat gekauft** das Haus.

*Hans **has** the house **bought**.* --> *Hans **has bought** the house.*

V *N P --> V P *N

```
ReorderFinVerb_hs =
        Ve{mark=hs}e{mark=vg_fiv},
       *Ne{mark=hs}a{mark~=vg_ptc;vg_inf},
        Pe{mark=hs}e{mark=vg_ptc}
   :      p(move=V->VPN).
```

Negation

- Negation
  DE: Hans kommt **nicht.** **--> Hans does not come.**
  *Hans comes **not.***

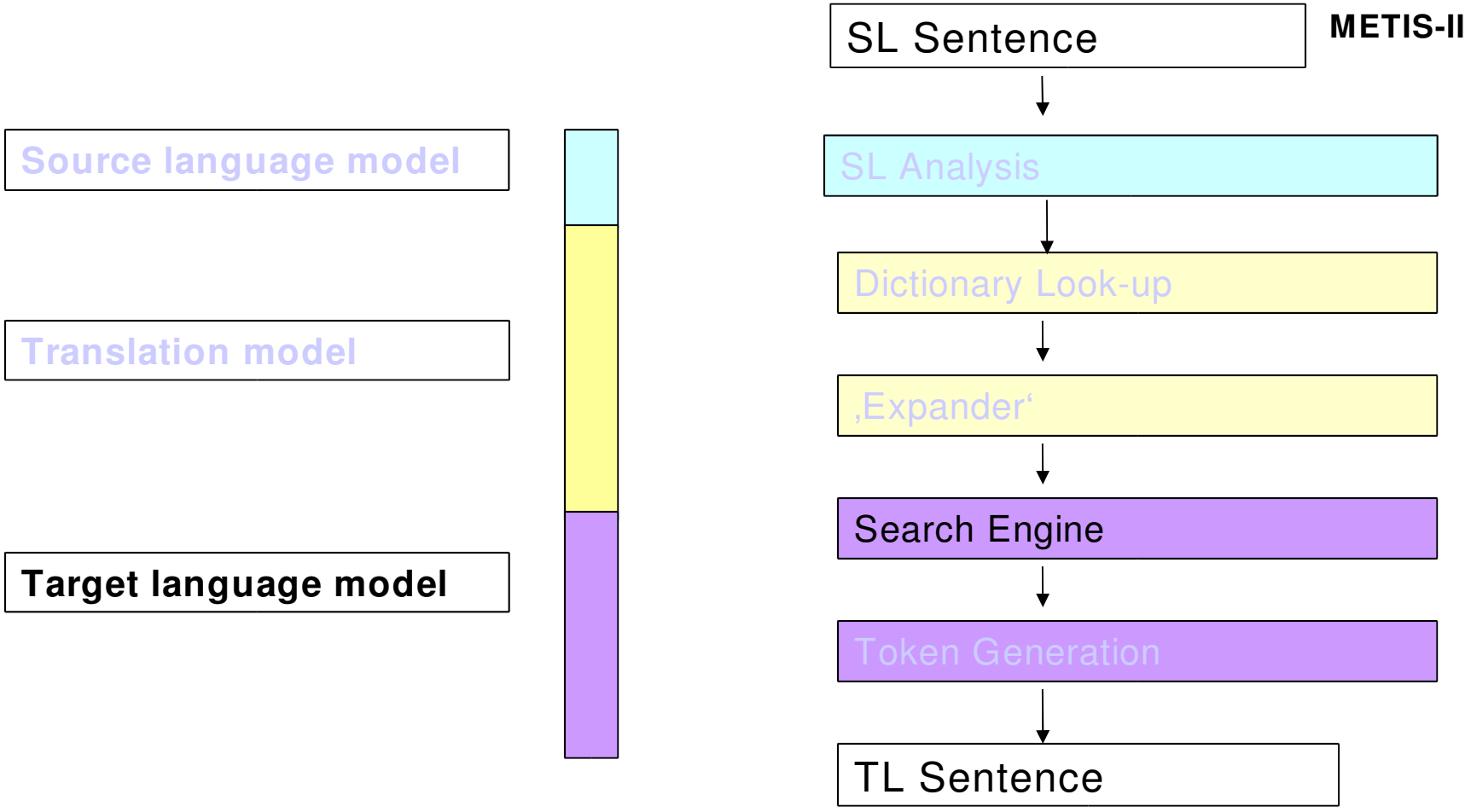
- Dictionary: nicht --> not | do not
- Expander Rule:

```
Negation_hs2 =
     Ae{mark=hs,mark=vg_fiv},
    *Be{mark=hs,lu~=nicht},
     Ne{mark=hs,lu=nicht}
   :   p(move=A->ANB).
```

AND/OR Graph for
Hans kommt nicht

- {lu=Hans,c=noun, wnr=1}
  @{c=noun}@{lu=hans,c=NP0}. .
  ,{lu=nicht,c=adv,wnr=3}
  @{c=verb}@{lu=do,c=VDZ},{lu=not,c=XX0}.
  , {c=adv}@{lu=not,c=XX0}..
  ,{lu=kommen,c=verb,wnr=2}
  @{c=verb}@{lu=come,c=VVB}.
  , {c=verb}@{lu=come,c=VVB},{lu=along,c=AVP}.
  , {c=verb}@{lu=come,c=VVB},{lu=off,c=AVP}.
  , {c=verb}@{lu=come,c=VVB},{lu=up,c=AVP}..
  .

TL Model: Search Engine



Search Engine: Scoring n-best Translations

METIS-II

- Beam-search algorithm (breadth first)
- Traverses AND/OR graph to score *n*-best Translations
- Heuristic Function :

$$\hat{e} = \operatorname{argmax} \sum_{m}^{M} w_m h_m(\cdot)$$

- h_i Feature Funktion
- w_i weigting
- Log-linear Combination of feature functions

Heuristic Function

METIS-II

- trained on BNC (10^8 words, 10^6 sentences)
- $LM(lem)$ Lemma Language Model (3-gram, 4-gram)
- $LM(tag)$ Tag Language Model (5-gram to 7-gram)
- $w(lem, tag)$ Lemma/tag co-ocurrence modell

$$\hat{e} = \text{argmax} \ \{w_1 * LM(tag) + w_2 * LM(lem) + w_3 * w(lem, tag)\}$$

lemma	tag	#	w(lem; tag)
tape-recorder	AJ0	3	1.003
tape-recorder	NN1	87	22.080
tape-recorder	NN2	13	3.512
tape-recorder	<*>	0	0.250

Search Engine Output

METIS-II

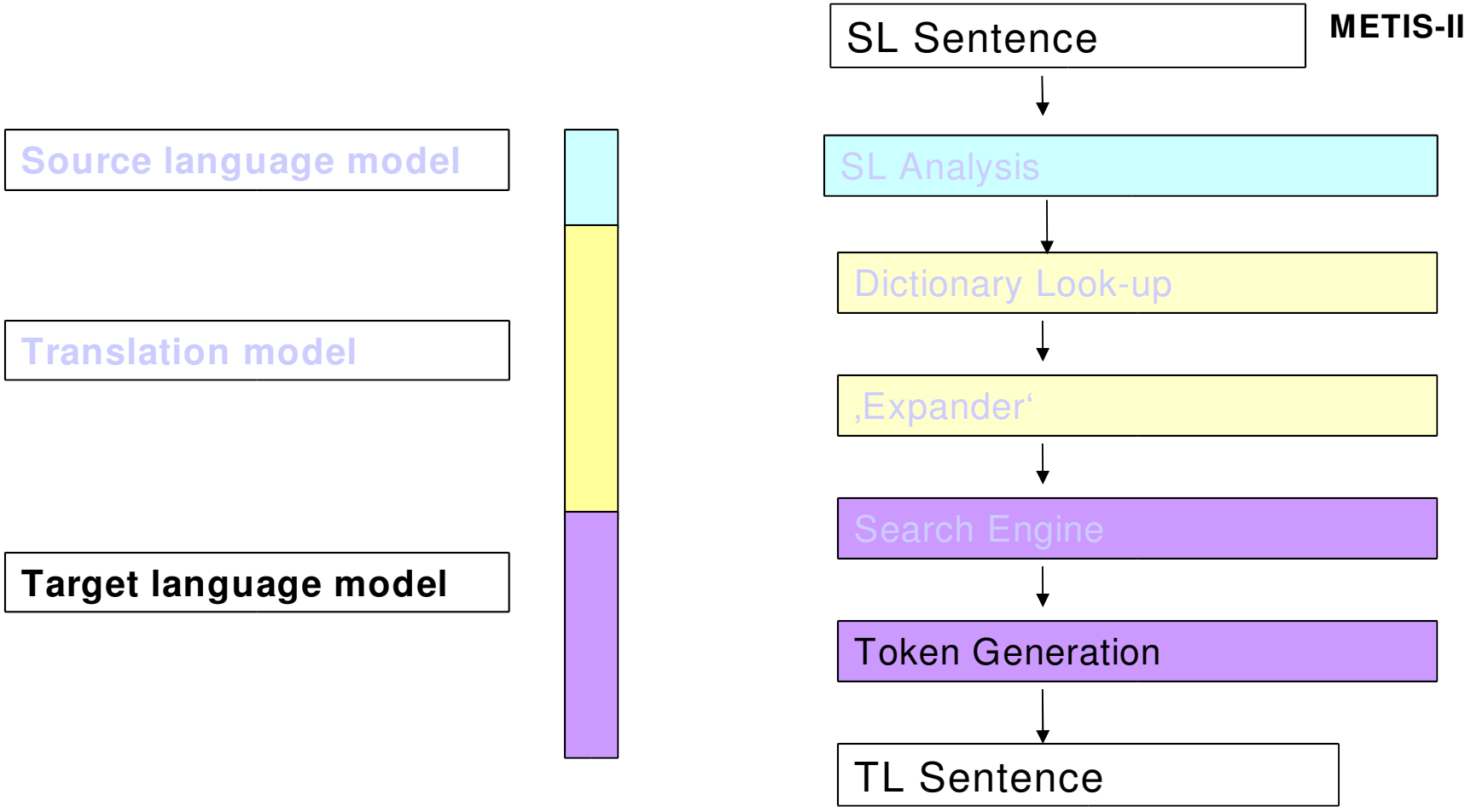
lemma, tag, #dictionary, expander rule:

<s id=3-0 lp="-9.227912">

the	AT0	146471	
company	NN1	268244	
is	VBD	604071	PermFinVerb_hs
buy	VVN	307263	PermFinVerb_hs
by	PRP	587268	PermFinVerb_hs
hans	NP0	265524	PermFinVerb_hs
.	PUN	367491	

</s>

TL Model: Token Generation



Reversible Lemmatiser and Token Generator

METIS-II

- Token Generator (for English):
  Lemma + Tag --> Token
  trained on reversible Lemmatiser:
  Token + Tag --> Lemma

- Reversible lemmatizer:
  - based on BNC Lemmatiser (10^8 words)
    - inflection rules are regular-expressions
  - augmented with additional tags (for reversibility)
    - 10 types of inflection rules for ADJ, NN2, VVG ...
    - ca 200 rules and ca. 1500 lexical exceptions

Reversible Lemmatiser

METIS-II

- Token + Tag <---> Lemma+Tag+O+IR

  Setting VVG <---> set VVG_f_4

- Lemma: normalised lemma
- Tag: BNC /CLAWS5 tag set
- O: orthographic properties of token
- IR: inflexion rule

- 100% reversibel: no loss of information
  --> but O and IR are not known when generating

Lemmatisation and Token Generation Rules

METIS-II

Lemmatisation: knowing token + tag:

- apply first matching rule

#	Tag	token suffix		lem. Suffix	Example token --> lemma
1	VVG	ffing	-->	ff	stuffing --> stuff
2	VVG	^(.{1,3}ll)ing	-->	$1	selling --> sell
3	VVG	ssing	-->	ss	kissing --> kiss

....

Token generation: knowing lemma + tag:

- guess lemmatisation rule
- apply inverse lemmatisation rule

Token Generation

METIS-II

- Generation while guessing inflexion rule:

  abort VVG ---> aborting VVG

- Method: guess inflexion rule from suffix of lemma.
  - Collect 27.000 suffixes from lemmatised BNC

Tag + suffix	# inflection rules	
VVG	28	unknown lemma suffix
VVG + t	5	
VVG + rt	2	
VVG + ort	2	
VVG + bort	1	deterministic token generation

- The longer the known lemma suffix the better the guess

Evaluation of Reversible Lemmatiser

METIS-II

- Lemmatiser:
  - 96,18% correct lemmas
  - incorrect mostly for closed class words (he, the, a, ...)

- Token Generator:
  - 99.5% correct reproduction of original token
    tested on 244,500 different wordforms
  - incorrect for writing variants:
    burned / burnt VVN --> burn VVN
    BNC british English: burned more likely

Conclusion

METIS-II

METIS-II German-to-English Basic Idea:

- First: use 'secure' symbolic resources:
  - generate partial translation hypotheses
  - store hypotheses in an AND/OR graph
- Then: use statistical resources:
  - rank best combination of partial translation hypotheses
  - integrate various global resources with feature functions

Main Components

METIS-II

- Lexicon:
  - basic translation equivalences
  - match phrases and discontinuous entries
  - overgeneration
- Expander:
  - structural adjustment
  - permute, insert, delete translation units
- Search engine:
  - rank translation hypotheses
  - use target language knowledge

Distribution of Information

- Search Engine vs. Lexicon

  stark <--> heavy,strong,large,big …

  Raucher <--> smoker

  or

  starker Raucher <--> heavy smoker

- Expander vs. Lexicon

  guerra <--> war

  civil <--> civil

  espagnol <--> spanish

  or

  guerra civil espagnol <--> spanish civil war

Evaluation

METIS-II

- Evaluation depends on:
  - Dictionary and performance of matching algorithm
  - Expander rules
  - Number and weights of feature functions

- Impact of modifications on BLEU score:
  - Changing weights of feature functions: BLEU scores from 1.6 to 1.8
  - Modifying expander rules: BLEU scores from 1.6 to 2.2

- Test set of 200 German sentences:

NIST	BLEU	lemma LM	tag LM
5.4801	0.1861	6M-n3	100K-n4
5.3004	0.2030	5M-n3	5M-n7

Future prospectives

METIS-II

- Enhancement of components:
  - Dictionary lookup (maintainance & matching)
  - Generation of discontinuous English fragments e.g. *give <sth> away, make <sth> easy*
  - Testing more feature functions (lexical weight)

- Dynamic Adaption to user needs:
  - explore automatised weighing strategies for:
    - Dictionary entries
    - Expander rules

THANK YOU