

Machine Translation

Symbolic Methods

Martin Kay

Stanford University and
The University of the Saarland

Elimination of
Exceptions
Special Cases

These things do not translate, though they may be involved in something that does

Decla

Prepositions
Moods

Morphographemic Abstraction

walking	=	walk	+	+ing
rubbing	=	rub	+	+ing
walks	=	walk	+	+s
tries	=	try	+	+s

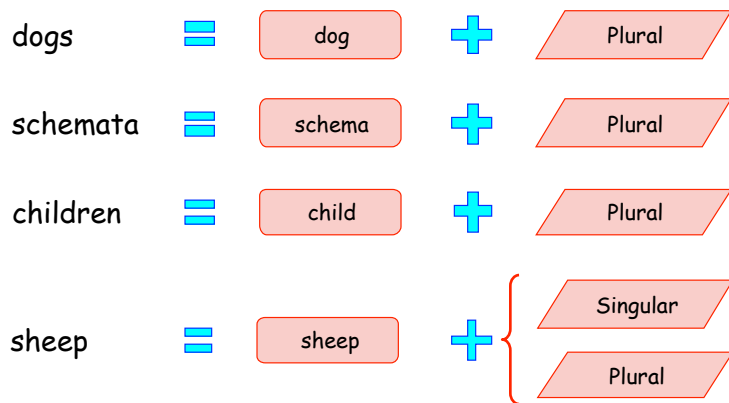
Spelling idiosyncracies no longer matter
no longer get in the way

Morphographemics

Kind	Kinder	Kindern
love	loves	loving
run	runs	running
manger	mange	mangeons
try	trying	tries
tie	tying	ties
medico		medici
arco		arche

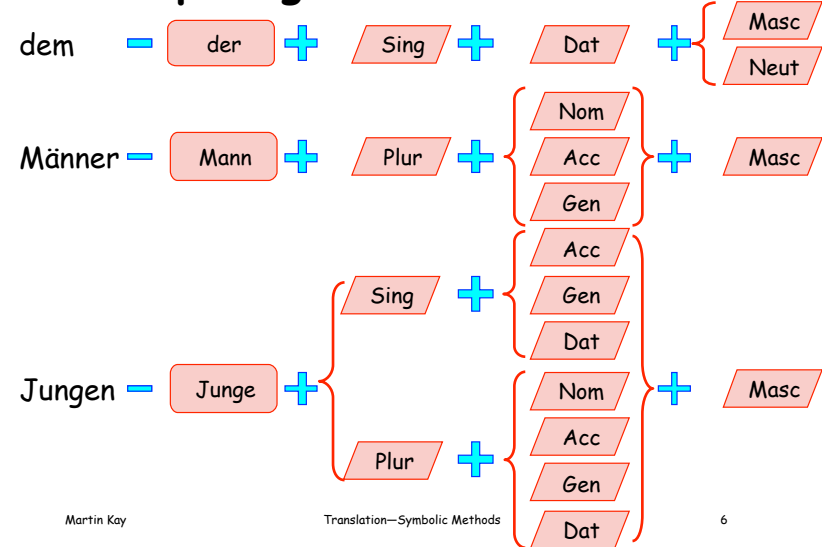
Diacritics

Morphological Abstraction

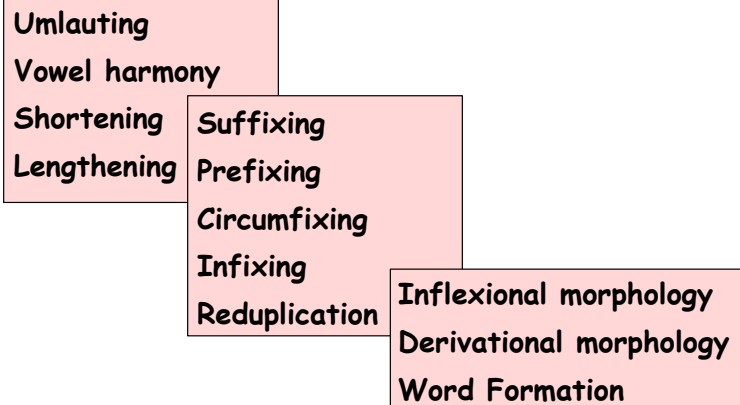


Paradigms and exceptions no longer matter

Morphological Abstraction



Word-level Processes



Morphemes vs. Structure

(Io) sono arrivato I arrived

(Loro) sono arrivati They (You) arrived

Il faut qu'il le fasse He must do it

Qu'il le fasse I hope he does it

Hans schwimmt gern Hans likes swimming

Sie können gern eins nehmen
Feel free to take one

What do you do for exercise?

I like swimming

I like to swim

I have to have this injection every week.
It is quite painful, so I like to have it done
on the weekend.

I have to have this injection every week.
It is quite painful, so I like having it done
on the weekend.

Syntactic Abstraction

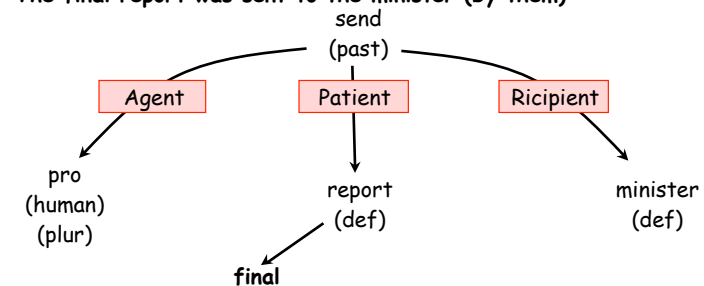
They sent the final report to the minister

They sent the minister the final report

The final report, they sent to the minister

To the minister they sent the final report

The final report was sent to the minister (by them)



Syntactic Abstraction

How much abstraction is enough/too much?

Information structure

John gave this perfect stranger a lot of money

John gave a lot of money to this perfect stranger

Broccoli, I cannot stand!

One thing I cannot stand is broccoli.

The more broccoli there is, the less I like it.

It is Ivan that caused all the trouble in the first place.

Topicalization

What does it mean in English/German?

Other Levels

His clever brother always stood in his light

Er stand immer im schatten seines klugen
Bruders

He will not be here until Monday

Er wird erst Montag da sein

Cela vous plait?

Do you like that?

Hans schwimmt gern

Hans likes swimming/to swim

*on Mary's bus by bus
in Mary's bus
in the bus
on the bus*

How did you get here?

Where did you leave your wallet?

Where is the fire extinguisher?

✓	?	✓	✓	✓
✓	✓	✓	?	x
✓	✓	✓	?	x

*In the chair next to me
On the chair next to me*

Where shall we put aunt Agatha?

Where shall I put this cushion?

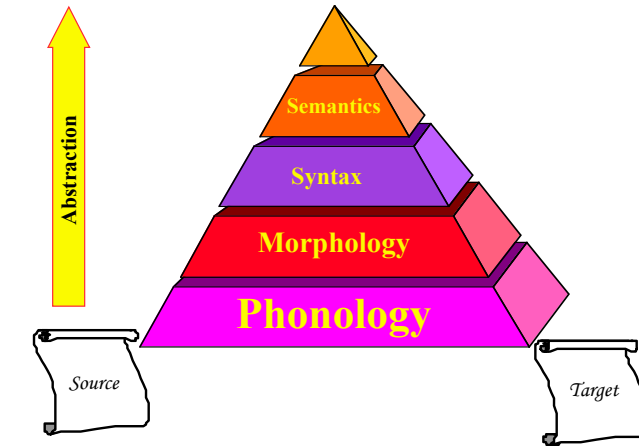
x	✓
✓	✓

Syntax? — Adjective order

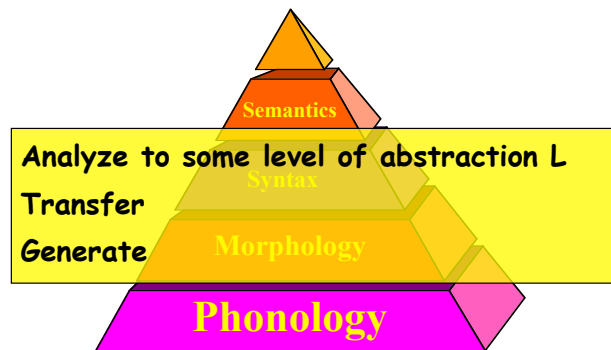
Opinion	Size	Age	Shape	Color	Origin	Material	Purpose	
Fine	big	old				wooden	storage	boxes
	little			blue	Mexican			model
Funny			round				meeting	room
					farm	vegetable		product

How to classify
organic
recursive
soft
running
... ?

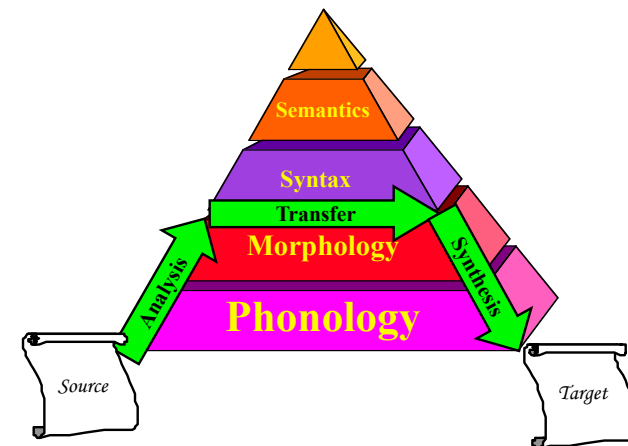
The Vauquois Triangle



The Transfer Approach



The Vauquois Triangle



Commercial Systems

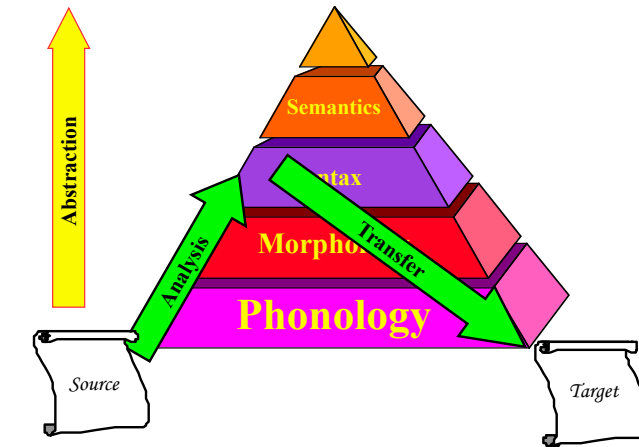
Do not follow the model closely:

- Levels of abstraction are
 - Not strongly separated
 - Are weakly formalized at best
- Generation Levels are largely eliminated

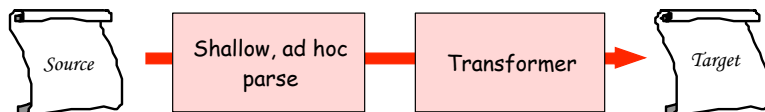
Commercial systems are almost entirely deterministic

Aim for speed

The Vauquois Triangle



The Standard Approach



Commercial Systems

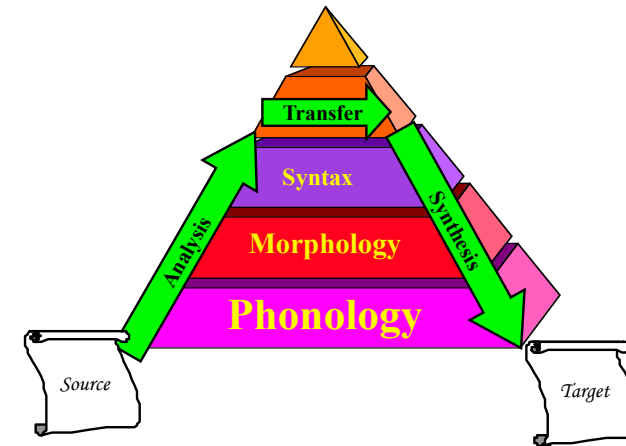
Rely on

- Tuning the lexicon to the domain
- Huge inventories of set phrases
- Selectional restrictions

Assessment of the Standard Approach

- Robust
- Can produce word salad
- Ad hoc and hard to maintain
- Bilingual and unidirectional

Academic Approaches



Orthography

Easy technology ~ finite-state

die	dies	dying	died
dye	dyes	dyeing	dyed
singe	singes	singeing	singed
develop	develops	developing	developed
stoop	stoops	stooping	stooped
enter	enters	entering	entered
bare	bares	baring	bared
hop	hops	hopping	hopped
travel	travels	traveling	traveled
travel	travels	travelling	travelled
humbug	humbugs	humberging	humberged
panic	panics	panicking	panicked
bus	buses	bussing	bussed
bus	buses	busing	bused
hoe	hoes	hoeing	hoed
pass	passes	passing	passed
buzz	buzzes	buzzing	buzzed

coax	coaxes	coaxing	coaxed
watch	watches	watching	watched
wash	washes	washing	washed
veto	vetoed	vetoing	vetoed
tie	ties	tying	tied
ski	skis	skiing	skied
play	plays	playing	played

English Morphographics

```

define sib          [j l s l x l z l s h l c h] ;
define consonant   [b l c l d l f l g l h l j l k l l l m l n l p l
                    q l r l s l t l v l w l x l y l z] ;

define vowel       [a l e l i l o l u] ;
define boundary    [.#. | % +];
define optional    [%? (->) 0] ;
define YtoIE       [ y -> i e || consonant _ EM alpha ] ;
define IEtoY       [ i e -> y || _ EM i ] ;
define Edeletion1  [ e -> 0 || vowel consonant _ EM vowel ] ;
define Edeletion2  [ e EM e -> EM e ] ;
define Einsertion  [ [ .. ] -> e || [sib l o] (diacritic) EM _ s EM ] ;
define gemination  [ b -> b b, c -> c k, d -> d d, f -> f f, g -> g g,
                    l -> l l, m -> m m, n -> n n, p -> p p, r -> r r,
                    s -> s s, t -> t t || vowel _ EM vowel ] ;

define DiacriticDeletion [ diacritic -> 0 ] ;
define BoundaryDeletion [ [ BM | EM ] -> 0 ] ;

```

```

define Word [[ preamble .o.
              optional .o.
              YtoIE .o.
              IEtoY .o.
              Einsertion .o.
              gemination .o.
              DiacriticDeletion .o.
              Edeletion1 .o.
              Edeletion2 .o.
              BoundaryDeletion] | 0 ];

```

Morphology

Prefix, suffix, infix, circumfix

Ablaut, umlaut, intercalation

agglutinating, polysynthetic languages

Compounding

Morphology

Generally finite-state

English Inflexion ~ easy, robust

Can be ambiguous, but not all that often

Irregular and suppletive forms

English Derivation ~ complex, fairly robust

Most people pretend it is not there

Occasional "syntactic" ambiguities: untiable, undoable.

Segmentation ambiguities: unionize

Overgeneration: redecomposablizationally

Others can be hard

Bantu, Finish, Sanskrit ...

What to do with Morphology?

- Type/token ratio
- POS Tag
- Shallow Syntax
 - NP Chunking
- Deep Syntax

Deep(?) Syntax

- Probabilistic Phrase structure/dependency grammar
- Dependency parsing
- LFG/HPSG/CCG ...

Deep Syntax

- Hugely ambiguous
 - *Gepard*: average ambiguity over a corpus of newspaper text (avg. 11.43 words): 78 readings
- Not robust
 - Language boundary is not well defined
 - Subcategorization
 - "Constructions"

Shallow Parsing

- Captures local phenomena at best.
- Fast — essentially finite-state
- Result may not be grammatical

Parsing with Fragments (LFG)

- A typical breakdown of parsing time of XLE components with the English grammar is
 - Morphology 1.6%,
 - Chart 5.8%
 - Unifier 92.6%.
- In the case of German, the typical time of XLE components is:
 - Morphology 22.5%,
 - Chart 3.5%
 - Unifier 74%

Transfer

Robust Parsing

- Any two words or phrases can form a phrase—at a cost.
- Arrange agenda items by cost
- Many different costs leads to poor performance because algorithm approximates breadth-first search.

Ambiguity

Time flies like an arrow

Fruit flies like a banana

Unplug the power cord from the wall outlet

Airport long term car park courtesy vehicle pickup point

I bought a car with four doors/dollars

Attach the end of the wire from the power supply of the unit to the red terminal on the panel at the back of the amplifier (1430 structures)

Connect pressure and return lines to pump

I just got back from Texas/Utah//Germany/Saudi Arabia. I had forgotten how good beer tastes.

Ich hatte vergessen, wie gut[es] Bier schmeckt.

His paper shows that smoking can cause cancer

- Order agenda by
 - Probability
 - Geometry—e.g. center embedding
 - Shallow processing—tags, chunks
 - Grammaticality
 - Known/unknown constructions

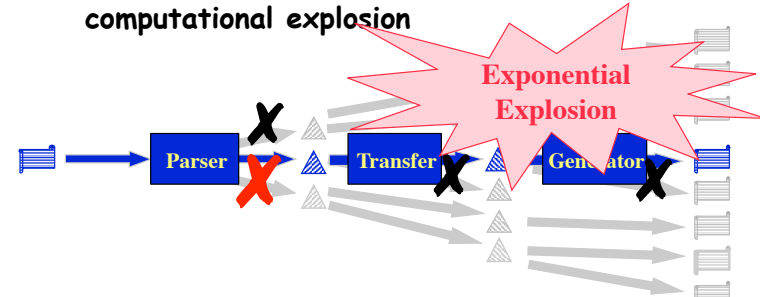
The Standard Approach

Separate modules for simplicity,
maintainability, reuse



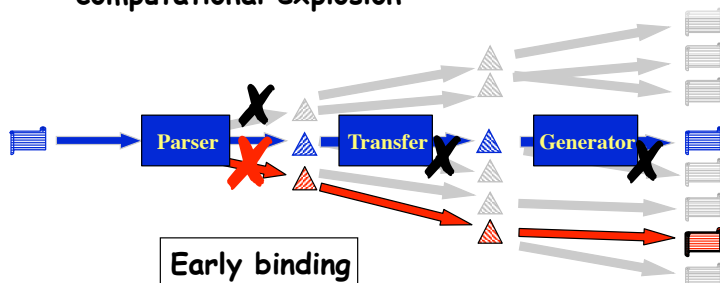
The Standard Approach

Separate modules for simplicity,
maintainability, reuse
Heuristic filters are applied early to avoid
computational explosion



The Standard Approach

Separate modules for simplicity,
maintainability, reuse
Heuristic filters are applied early to avoid
computational explosion



Academic Approaches

- More abstraction — appeal to AI
- Equal weight to analysis and generation
- Formalization
- Avoid early binding

Academic Approaches

Problems

Time

Robustness

Ambiguity

Linguistics

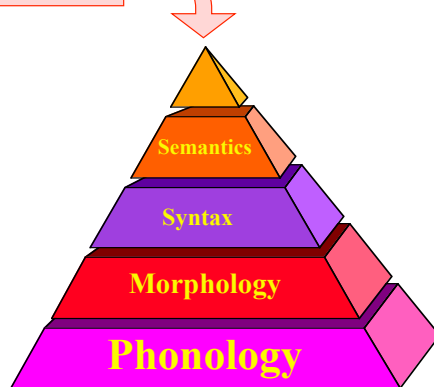
Can identify

But not resolve

Ambiguity

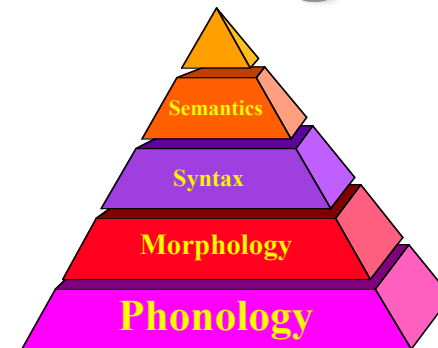
The Vauquois Triangle

What is this?



The Vauquois Triangle

Interlingua



If you abstract enough

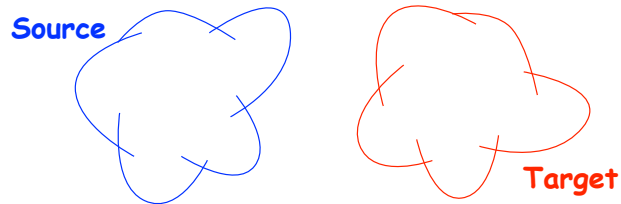
You will be left with Pure Thought

OK. So what is wrong with that?

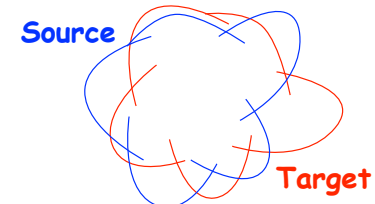
Interlingua must

- Represent whatever any language can represent, even if it will often be lost in translation.
- Problems of (non)overlap in the semantic grid.

- The power of natural language lies in the fact that it can be used casually. It neither requires, nor admits, precision (in things that matter).



- The power of natural language lies in the fact that it can be used casually. It neither requires, nor admits, precision (in things that matter).



- **(Weighted) Synchronous grammar**
 - Syntax-directed transduction grammar
 - Syntax directed translation schemata

Selectional Restrictions

The dog ate the meat

Der Hund hat das Fleisch gefressen

I know that he read the book

Je sais qu'il a lu le livre

I know the book

Je connais le livre

Minimal Recursion Semantics

- **Underspecification**
 - Quantification
 - Nesting

Nesting

Fierce black cat

$\lambda x[\text{fierce}(x) \wedge (\text{black}(x) \wedge \text{cat}(x))]$

Nesting

Fierce black cat

$\lambda x[\text{fierce}(x) \wedge (\text{black}(x) \wedge \text{cat}(x))]$

$\lambda x[\text{gato}(x) \wedge (\text{negro}(x) \wedge \text{feroz}(x))]$

Nesting

Fierce black cat

$\lambda x[\text{fierce}(x) \wedge (\text{black}(x) \wedge \text{cat}(x))]$

$\lambda x[\text{gato}(x) \wedge (\text{negro}(x) \wedge \text{feroz}(x))]$

gato negro y feroz

Nesting

Fierce black cat

$\lambda x[\text{fierce}(x) \wedge (\text{black}(x) \wedge \text{cat}(x))]$

$\lambda x[\text{gato}(x) \wedge (\text{negro}(x) \wedge \text{feroz}(x))]$

gato negro y feroz

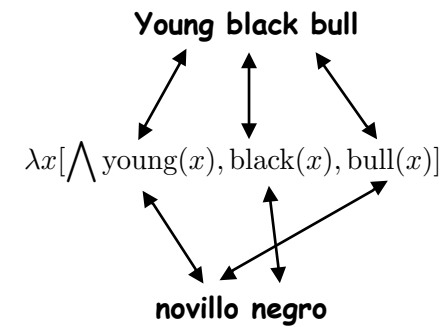
$\lambda x[\text{cat}(x) \wedge (\text{black}(x) \wedge \text{fierce}(x))]$

- **Try All Equivalent Forms**
- **Undecidable!**
- **Tune grammars to one another**
- **Too difficult**
- **Only good for on language pair**

- **Shake and Bake**
- **Try all possible orderings of words**
- **Synchronous grammar**
- **Must be lexicalized**

$\lambda x[\bigwedge \text{cat}(x), \text{black}(x), \text{fierce}(x)]$

$\lambda x[\bigwedge \text{black}(x), \text{cat}(x), \text{fierce}(x)]$



• **Preserve ambiguity of**

- Scope
- PP attachment

Neo-Davidsonian semantics — with event variables.

No spurious scoping distinctions

on(in(run(Kim), Foothills Park), Monday)

on(in(run(Kim), Monday), Foothills Park)

on(e , Monday) \wedge run(e , Kim) \wedge in(e , Foothills Park)

Every useful dictionary is big

every(x , dictionary(x) \wedge useful(x), big(x))

\bigwedge [every(x), dictionary(x), useful(x), big(x)]

Not enough information to reconstruct scope

Handles

\bigwedge [every(x , 1, 2), dictionary₁(x), useful₁(x), big₂(x)]

Every dog chased some cat

\bigwedge [every₁(x , 3, n), dog₃(x), cat₇(y), some₅(y , 7, m), chase₄(e , x , y)]

\bigwedge [every₁(x , 3, 4), dog₃(x), cat₇(y), some₅(y , 7, 1), chase₄(e , x , y)]

-- wide scope: some

\bigwedge [every₁(x , 3, 5), dog₃(x), cat₇(y), some₅(y , 7, 4), chase₄(e , x , y)]

-- wide scope: every

unscoped

$$\left[\begin{array}{l} \text{HANDEL} \quad \boxed{1} \vee \boxed{5} \\ \text{INDEX} \quad \boxed{8} \\ \text{LISZT} \left\langle \begin{array}{l} \text{HANDEL} \quad \boxed{1} \\ \text{BV} \quad \boxed{2} \\ \text{every_rel} \text{ RESTRICTION} \quad \boxed{3} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{HANDEL} \quad \boxed{3} \\ \text{INST} \quad \boxed{2} \\ \text{dog_rel} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{HANDEL} \quad \boxed{5} \\ \text{BV} \quad \boxed{6} \\ \text{some_rel} \text{ RESTRICTION} \quad \boxed{7} \end{array} \right\rangle, \\ \left\langle \begin{array}{l} \text{HANDEL} \quad \boxed{7} \\ \text{INST} \quad \boxed{6} \\ \text{cat_rel} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{HANDEL} \quad \boxed{1} \\ \text{EVENT} \quad \boxed{8} \\ \text{ACT} \quad \boxed{2} \\ \text{UND} \quad \boxed{6} \\ \text{chase_rel} \end{array} \right\rangle \end{array} \right]$$

Martin Kay Translation—Symbolic Methods 69

- Kim swam across the river
- Kim cruzó el río nadando
- Kim crossed the river (while) swimming

nadando:

$$\left[\begin{array}{l} \text{HANDEL} \quad \boxed{a1} \\ \text{INDEX} \quad \boxed{c2} \\ \text{LISZT} \left\langle \begin{array}{l} \text{HANDEL} \quad \boxed{a1} \\ \text{ARG1} \quad \boxed{c2} \\ \text{while_rel} \text{ ARG2} \quad \boxed{c1} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{HANDEL} \quad \boxed{a1} \\ \text{EVENT} \quad \boxed{c2} \\ \text{ACT} \quad \boxed{c1} \\ \text{nadar_rel} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{HANDEL} \quad \boxed{a1} \\ \text{EVENT} \quad \boxed{c1} \end{array} \right\rangle \end{array} \right]$$

Martin Kay Translation—Symbolic Methods 70

$$\left\langle \begin{array}{l} \text{EVENT} \quad \boxed{c1} \\ \text{ACT} \quad \boxed{c1} \\ \text{swim_rel} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{ARG1} \quad \boxed{c1} \\ \text{path_rel} \end{array} \right\rangle \Leftrightarrow \left\langle \begin{array}{l} \text{EVENT} \quad \boxed{c2} \\ \text{ACT} \quad \boxed{c1} \\ \text{nadar_rel} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{ARG1} \quad \boxed{c2} \\ \text{ARG2} \quad \boxed{c1} \\ \text{while_rel} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{EVENT} \quad \boxed{c1} \\ \text{ACT} \quad \boxed{c1} \\ \text{move_path_rel} \end{array} \right\rangle$$

$$\left\langle \begin{array}{l} \text{ARG1} \quad \boxed{c1} \\ \text{ARG2} \quad \boxed{c2} \\ \text{across_rel} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{EVENT} \quad \boxed{c1} \\ \text{ACT} \quad \boxed{c1} \\ \text{move_rel} \end{array} \right\rangle \Leftrightarrow \left\langle \begin{array}{l} \text{EVENT} \quad \boxed{c1} \\ \text{ACT} \quad \boxed{c1} \\ \text{PATH} \quad \boxed{c2} \\ \text{crosser_rel} \end{array} \right\rangle$$

In combination:

$$\left\langle \begin{array}{l} \text{EVENT} \quad \boxed{c1} \\ \text{ACT} \quad \boxed{c1} \\ \text{swim_rel} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{ARG1} \quad \boxed{c1} \\ \text{ARG2} \quad \boxed{c2} \\ \text{across_rel} \end{array} \right\rangle \Leftrightarrow \left\langle \begin{array}{l} \text{EVENT} \quad \boxed{c2} \\ \text{ACT} \quad \boxed{c1} \\ \text{nadar_rel} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{ARG1} \quad \boxed{c2} \\ \text{ARG2} \quad \boxed{c1} \\ \text{while_rel} \end{array} \right\rangle, \left\langle \begin{array}{l} \text{EVENT} \quad \boxed{c1} \\ \text{ACT} \quad \boxed{c1} \\ \text{PATH} \quad \boxed{c2} \\ \text{crosser_rel} \end{array} \right\rangle$$

Martin Kay Translation—Symbolic Methods 71

Problems

- Too many solutions
 - Packing?
- Robustness (= too few solutions!)
 - Often lexical gaps. Subcategorization
- Speed
- Shallow methods—fast and more robust
 - POS tagging
 - NP chunking

Martin Kay Translation—Symbolic Methods 72

Done!