6.863J Natural Language Processing
Lecture 20: Machine translation 4

Robert C. Berwick

The Menu Bar

- Administrivia:
  - final projects –
  - *Agenda:*
  - Combining statistics with language knowledge in MT
  - MT – the statistical approach (the "low road")
    - Evaluation
    - Where does it go wrong? Beyond the "talking dog"
  - MT – Middleuropa ground
    - Transfer Approach: using syntax to help

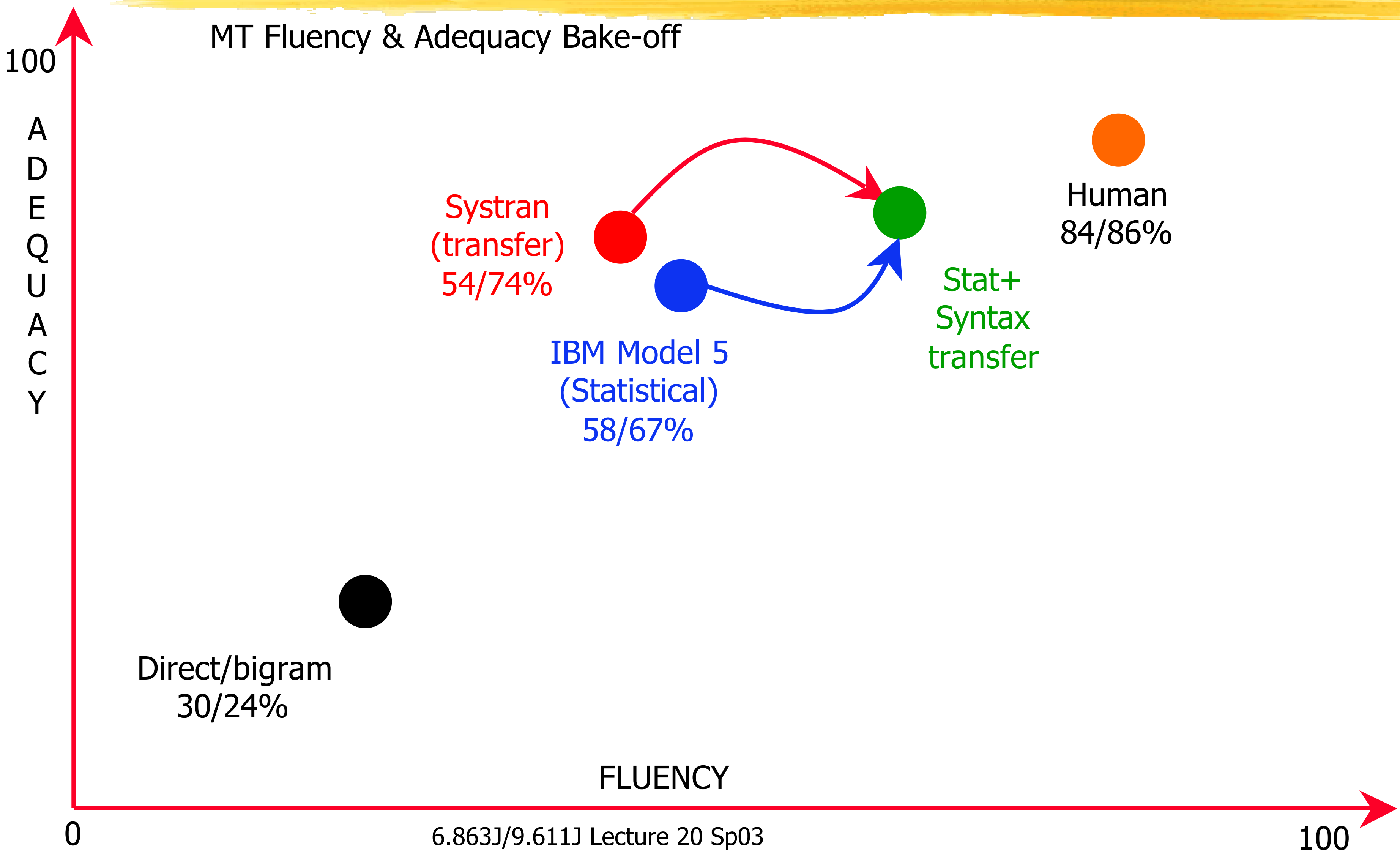
  How to combine w/ statistical information

Can we attain the Holy Grail?

How well does stat MT do?

- What happens if the sentence is <u>already seen</u> (part of training pair)?
- Then the system works <u>just as hard</u>
- Remembrance of translations past…?
- We get "only" 60% accuracy (but better than Systran…)
- Let's see how to <u>improve</u> this by adding knowledge re syntax
- Probably even better to add knowledge re semantics… as we shall see

The game plan to get better



Problemos

- F in: L'atmosphère de la Terre rend un peu myopes mêmes les meilleurs de leur télèscopes
- E out: The atmosphere of the Earth returns a little myopes same the best ones of their telescopes
- (Systran): The atmosphere of the Earth makes a little short-sighted same the best of their télèscopes
- (Better) The earth's atmosphere makes even the best of their telescopes a little 'near sighted'
- Why?

Let's take a look at some results…

Should

should

f	t(f|e)	phi	(phi|e)
devrait	0.330	1	0.649
Devraient	0.123	0	0.336
devrions	0.109	2	0.014
faudrait	0.073		
faut	0.058		
doit	0.058		
aurait	0.041		
doivent	0.024		
devons	0.017		
devrais	0.013		

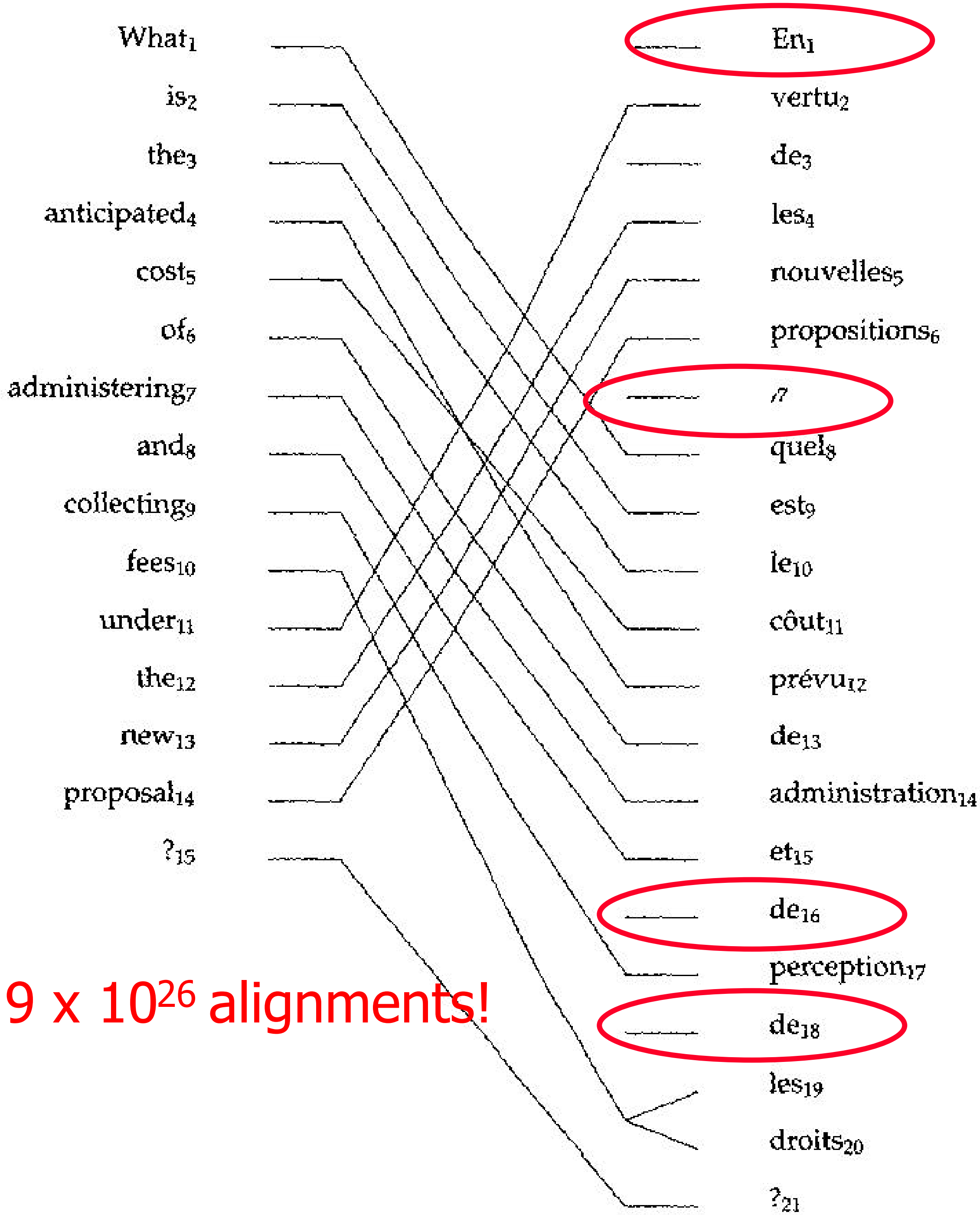
What about…

- In French, what is worth saying is worth saying in many different ways
- He is nodding:
  - Il fait signe qui oui
  - Il fait un signe de la tête
  - Il fait un signe de tête affirmatif
  - Il hoche la tête affirmativement

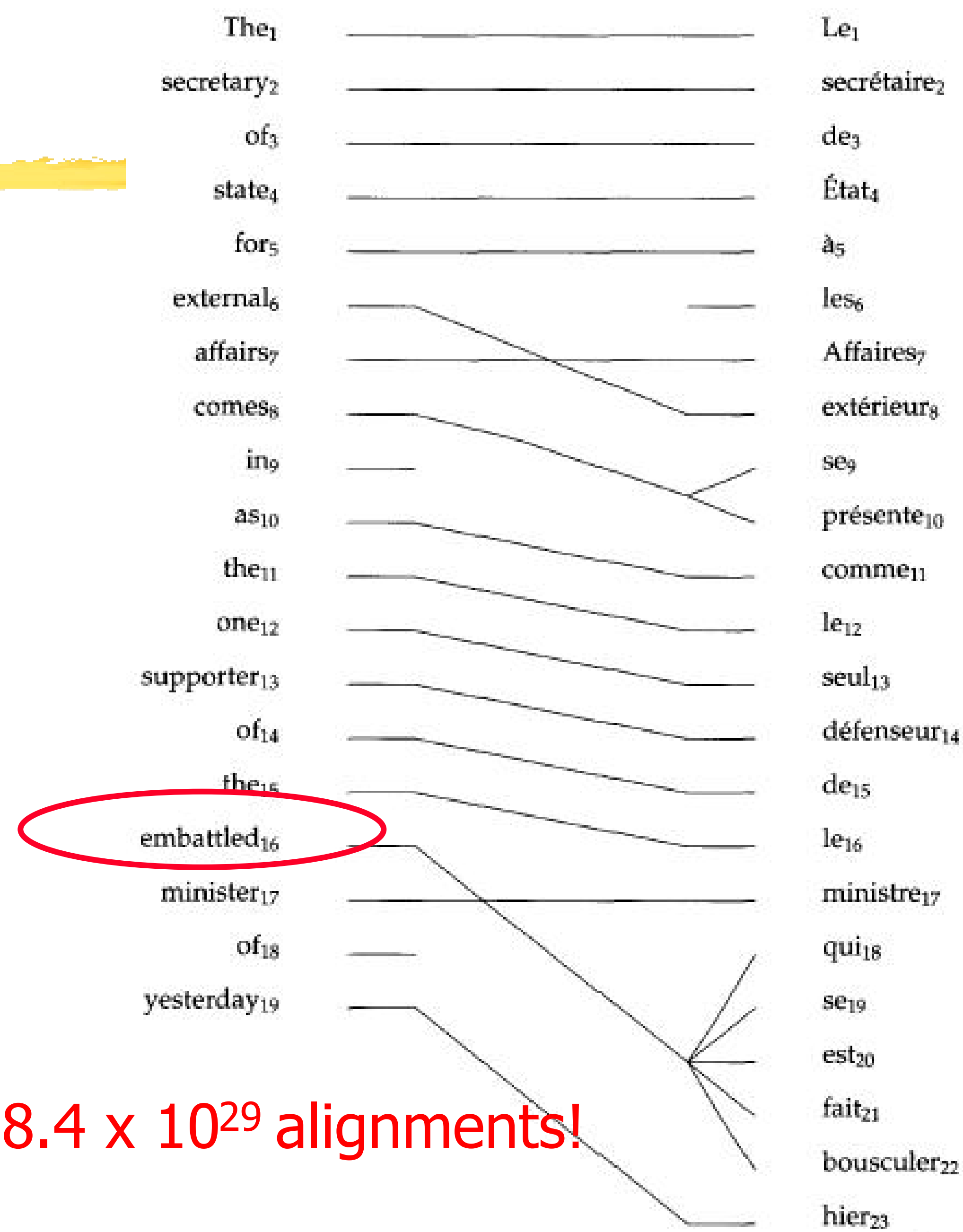
Nodding hill…

nodding

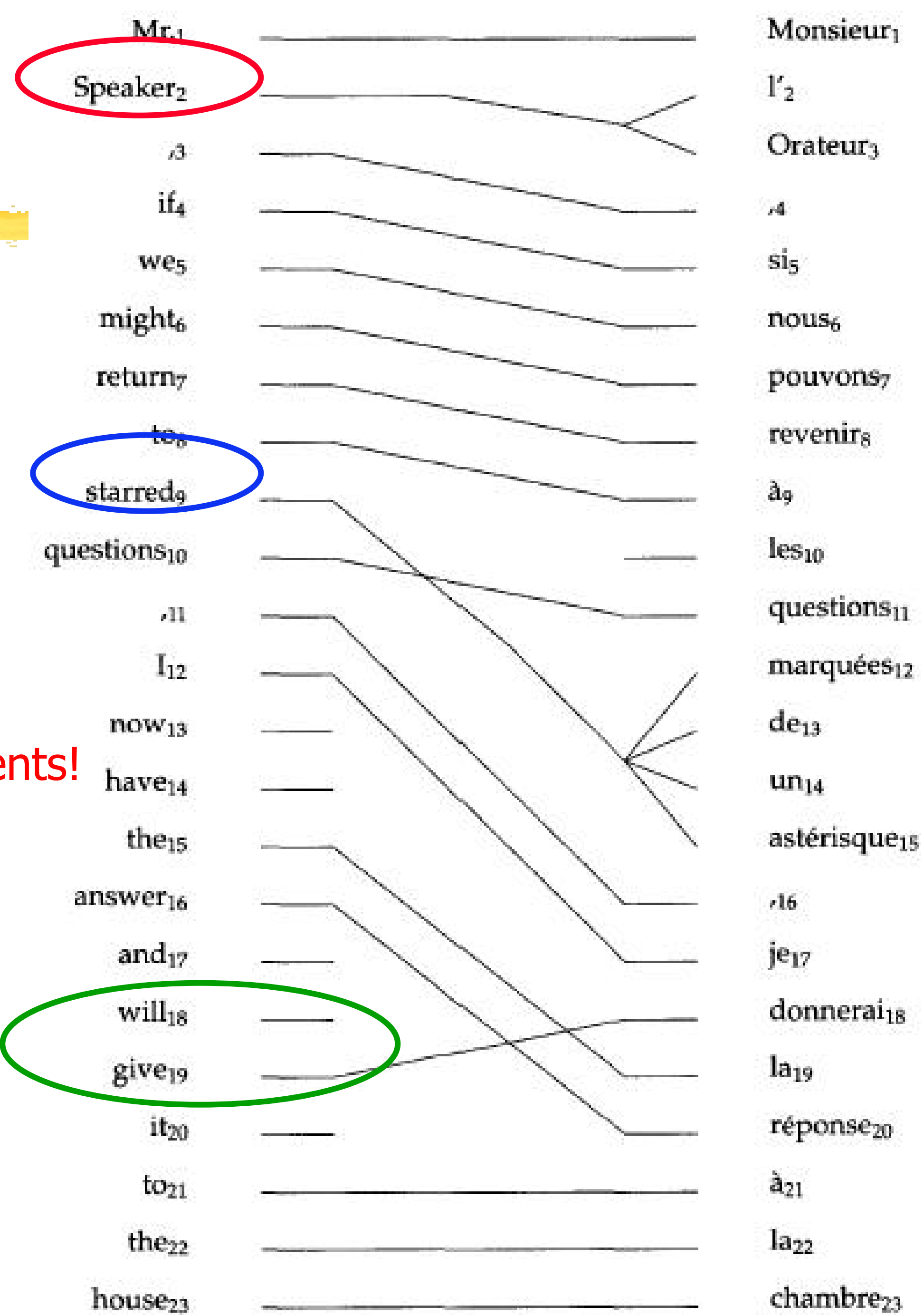
f	t(f|e)	phi	n(phi | e)
signe	0.164	4	0.342
la	0.123	3	0.293
tête	0.097	2	0.167
oui	0.086	1	0.163
fait	0.073	0	0.023
que	0.073		
hoche	0.054		
hocher	0.048		
faire	0.030		
me	0.024		
approuve	0.019		
qui	0.019		
un	0.012		
faites	0.011		



Best of 1.9 x 10^{26} alignments!



Best of 8.4 x 10^{29} alignments!



5.6 x 10^{31} alignments!

Morals? ¿Moralejas? ? ? ? ? .

- Always works hard – even if the input sentence is one of the training examples
- Ignores morphology – so what happens?
- Ignores phrasal chunks – can we include this? (Do we?)…
- Can we include syntax and semantics?
- (why not?)

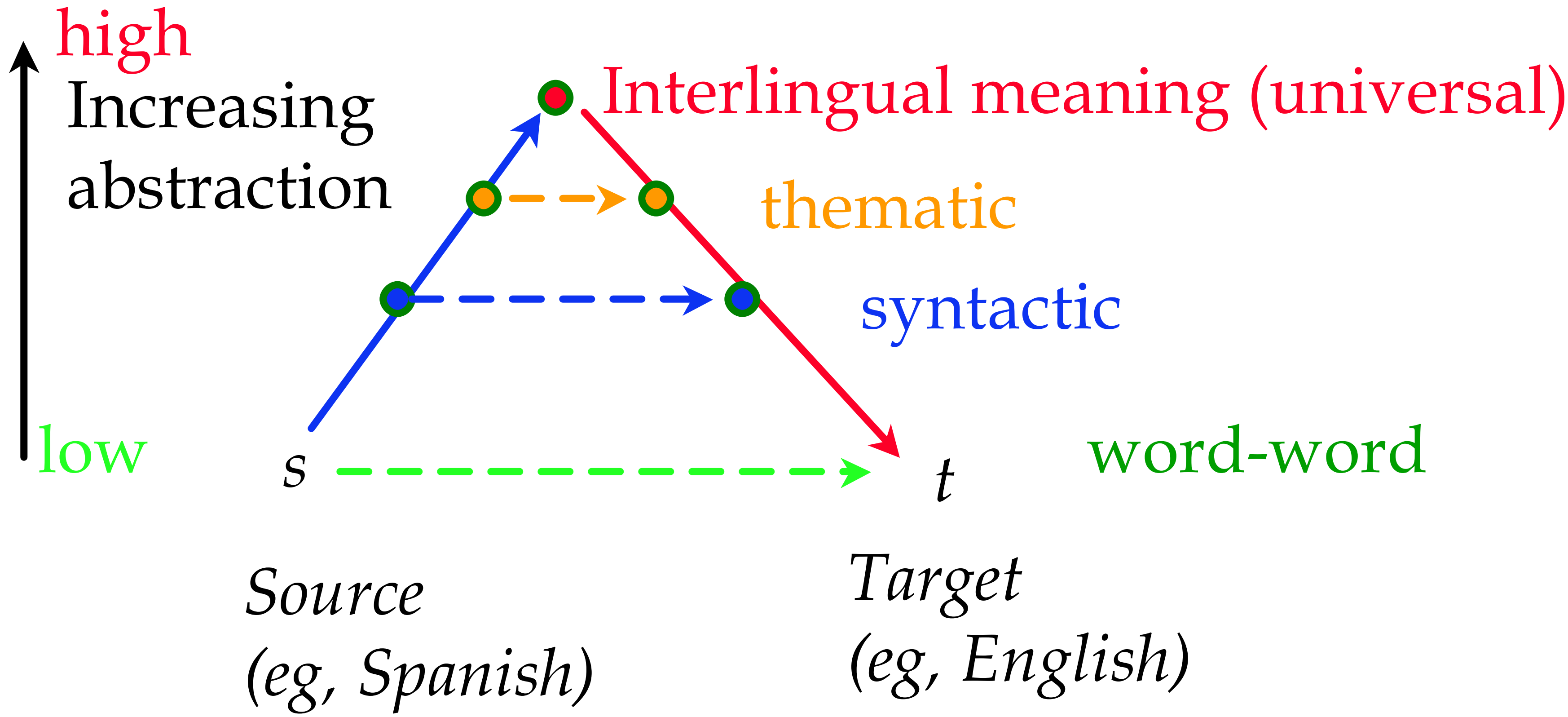
Other languages…

- Aligning corpus – a cottage industry
  - Instruction Manuals
  - Hong Kong Legislation - Hansards
  - Macao Legislation
  - Canadian Parliament Hansards
  - United Nations Reports
  - Official Journal
    of the European Communities

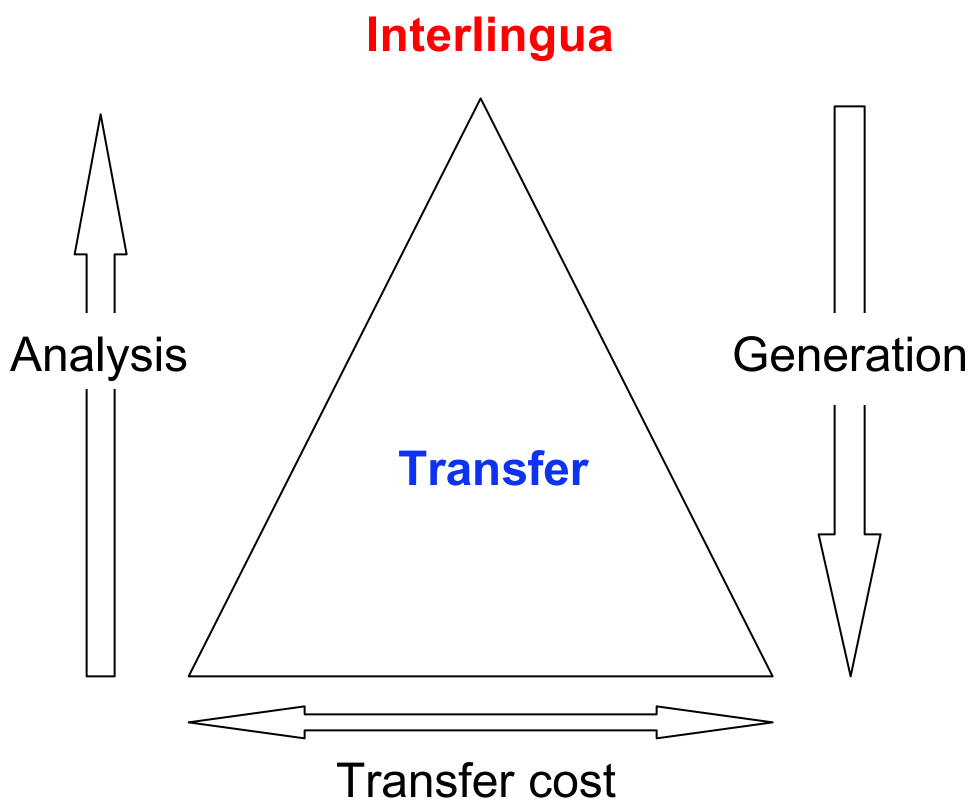
How can we do better?

- Systran: <u>transfer</u> approach
- Q: What's that?
- A: transfer <u>rules</u> for little bits of syntax
- Then: combine these rules with the statistical method
  - Even doing this <u>a little</u> will improve us to about 65%
  - Gung ho – we can get to 70%
  - Can we get to the magic number?

The golden (Bermuda?) triangle



The Bermuda triangle revisited



Vauquois Triangle

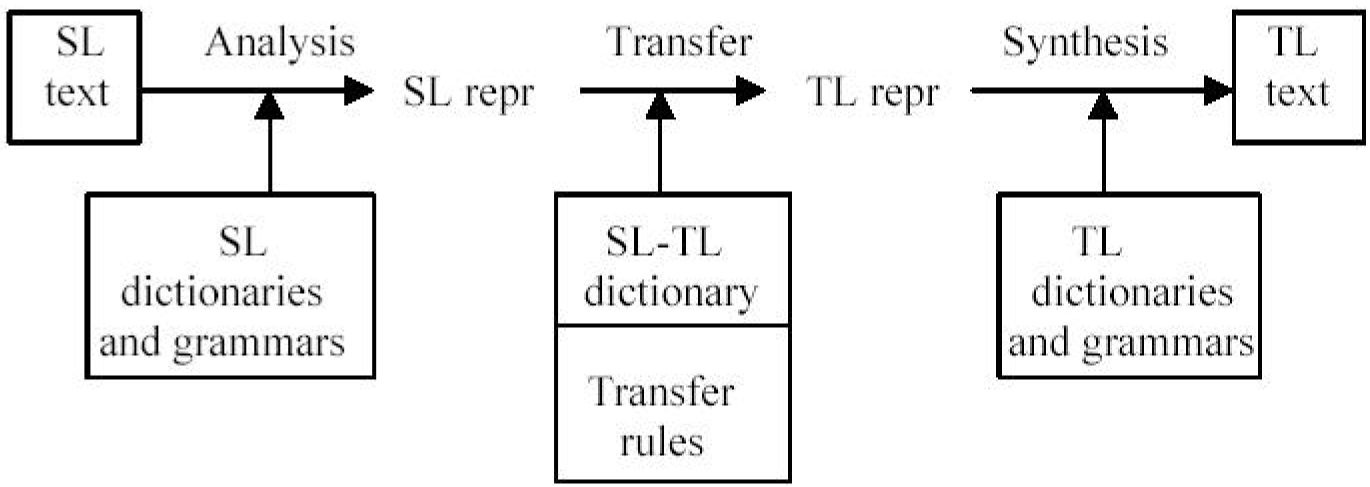
Transfer station

- Transfer: Contrasts are fundamental to translation. Statements in one theory (source language) are mapped into statements in another theory (target language)
- Interlingua: Meanings are language independent and can be encoded. They are extracted from Source sentences and rendered as Target sentences.

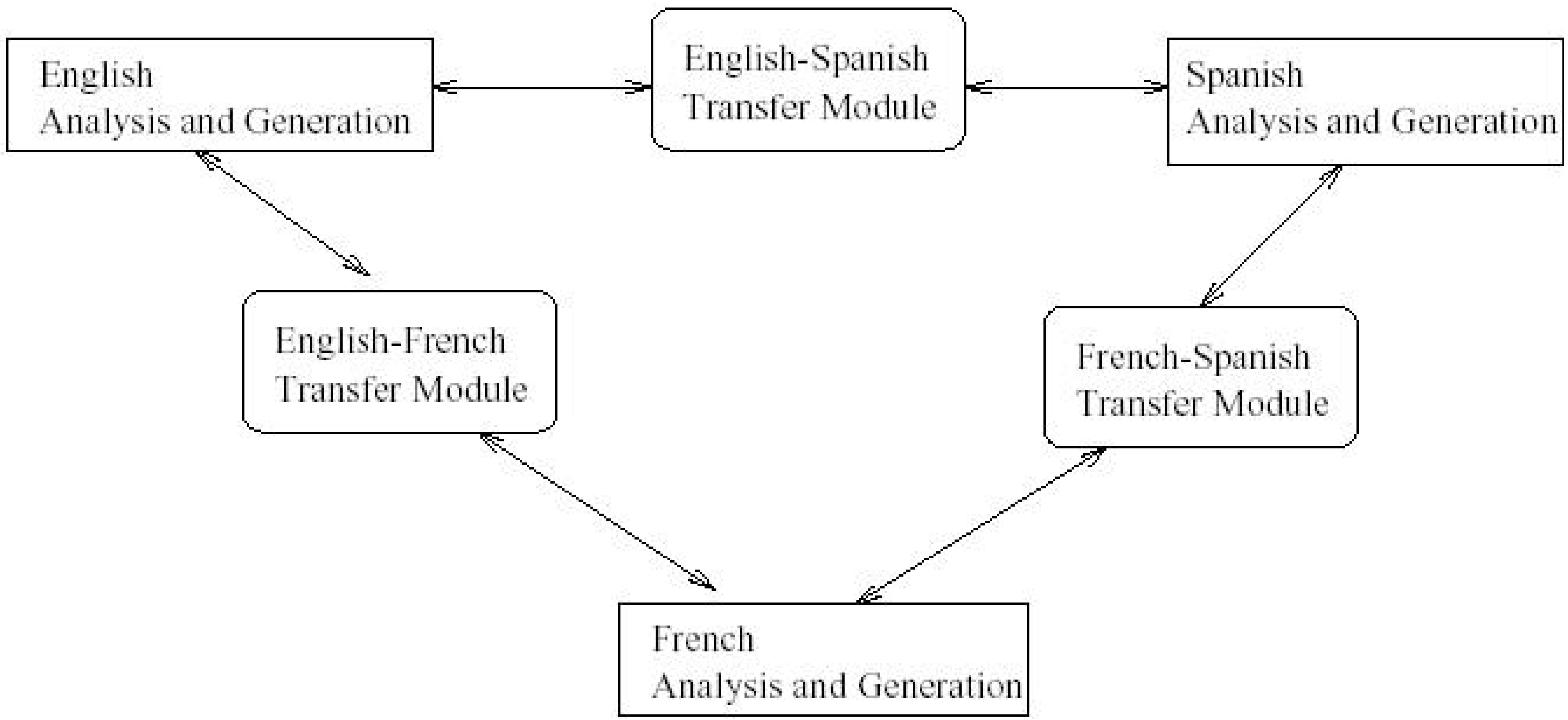
Transfer approach

- Analysis using a morphological analyser, parser and a grammar
- Depending on approach, grammar must build syntactic and/or semantic representation
- Transfer: mapping between S and T
- Generation using grammar and morphological synthesizer (from analysis?)

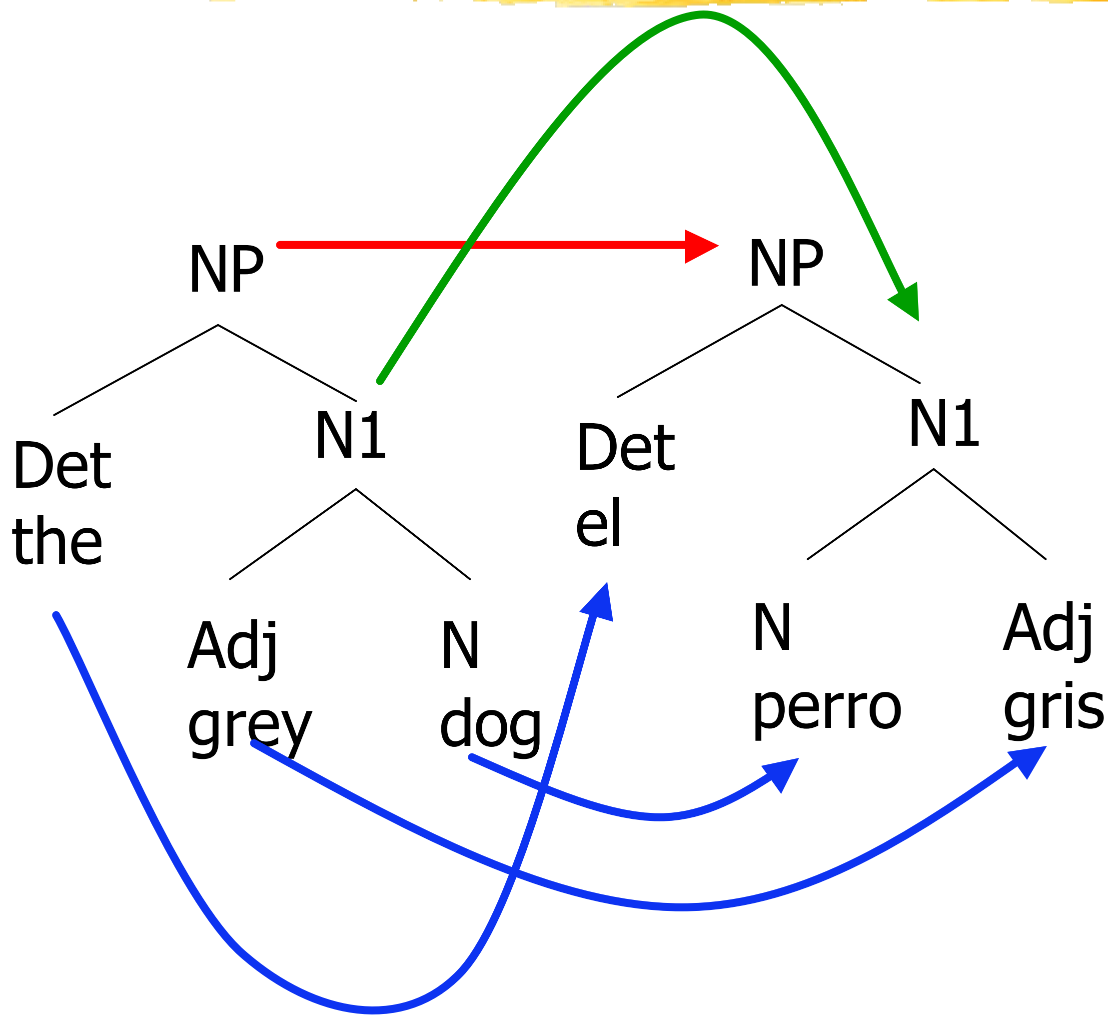
Transfer system: 2 languages



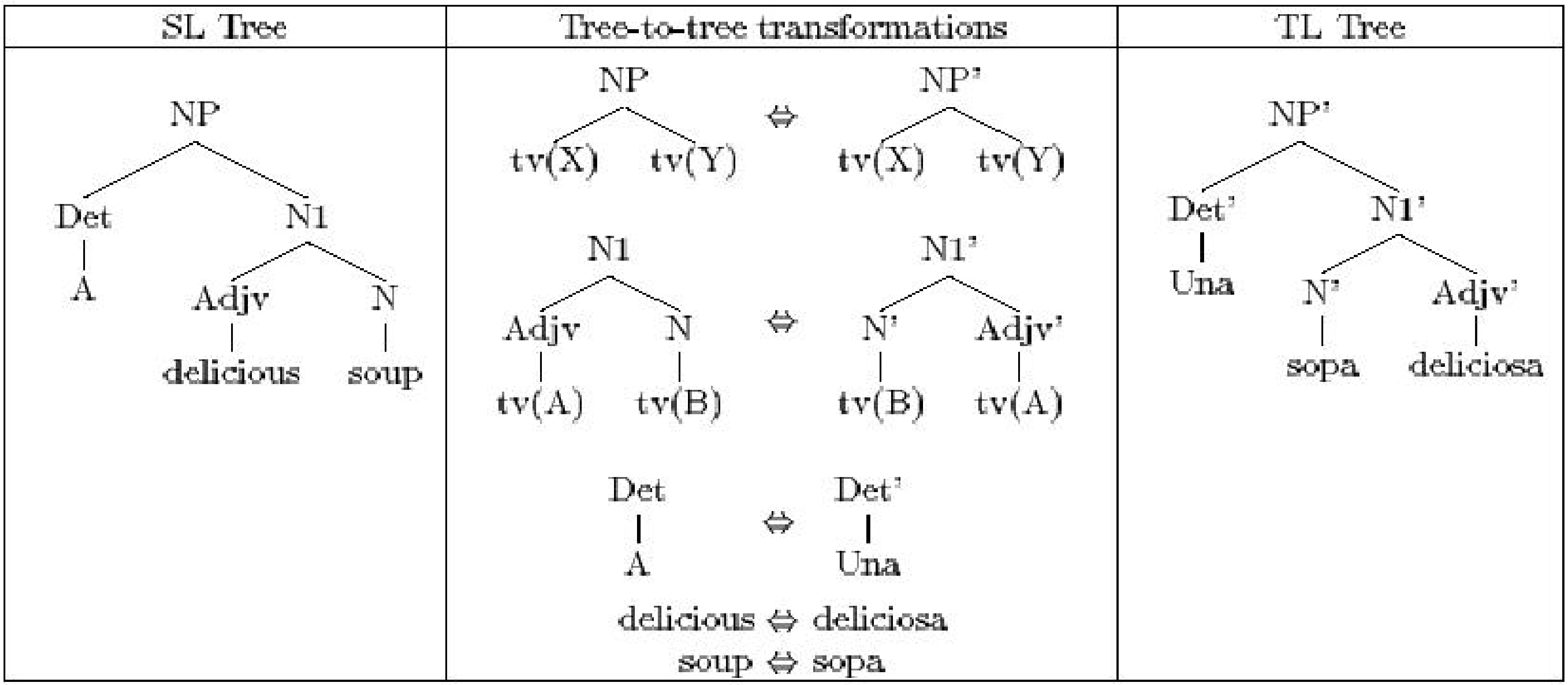
Transfer – multiple languages



Syntactic Transfer



Syntactic transfer



5 transfer rules: 3 syntax, 2 lexical

Syntactic transfer

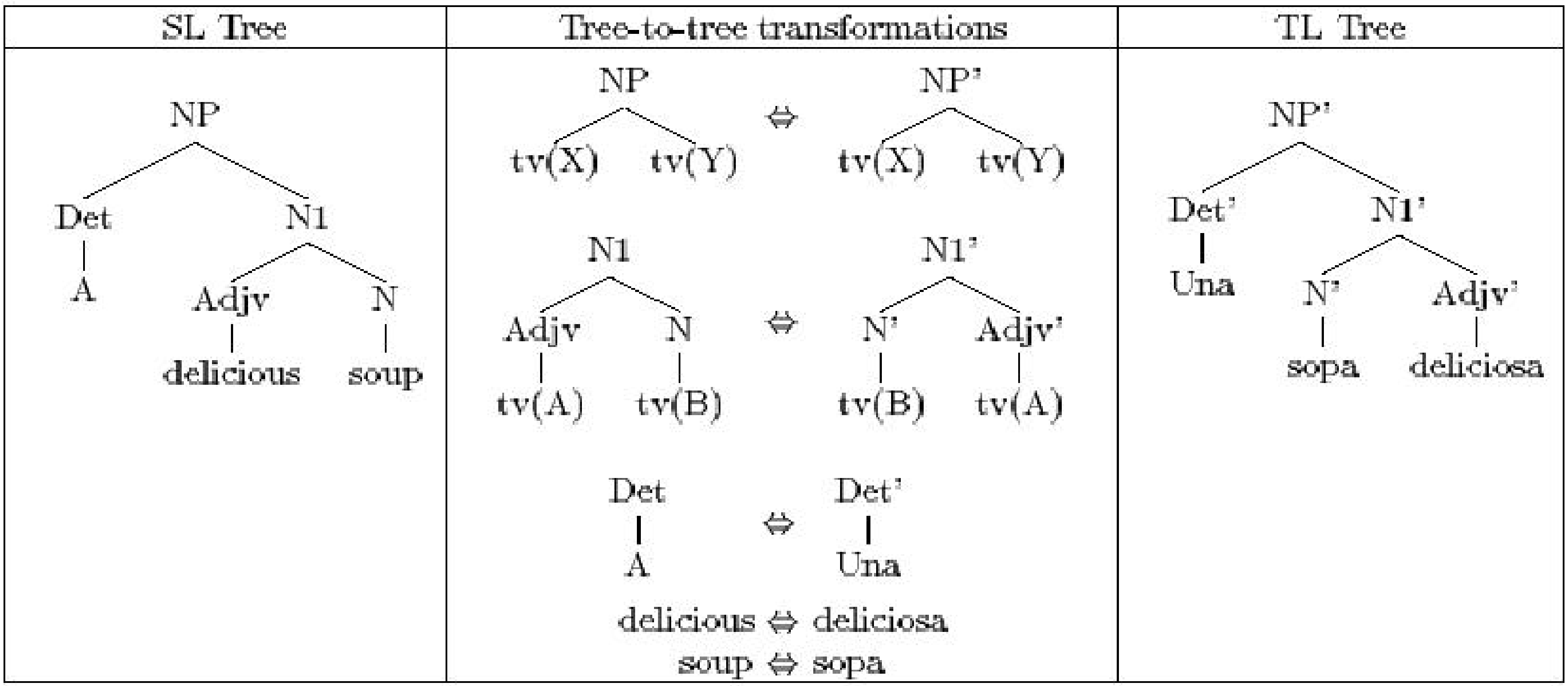
- Maps trees to trees
- No need for ‘generation’ except morphology
- Method: top-down recursive, non-deterministic match of transfer rules (where tv is a variable) against tree in source language
- Output is tree in target language (w/o word morphology)

Simple syntactic transfer example

- Rules (English-Spanish) – 3 in previous example
- 1 for NP NP; 1 for N1 N1'; one for Det Det
- Lexical correspondences

- Suppose input is as in preceding example – trace through matching

Syntactic transfer

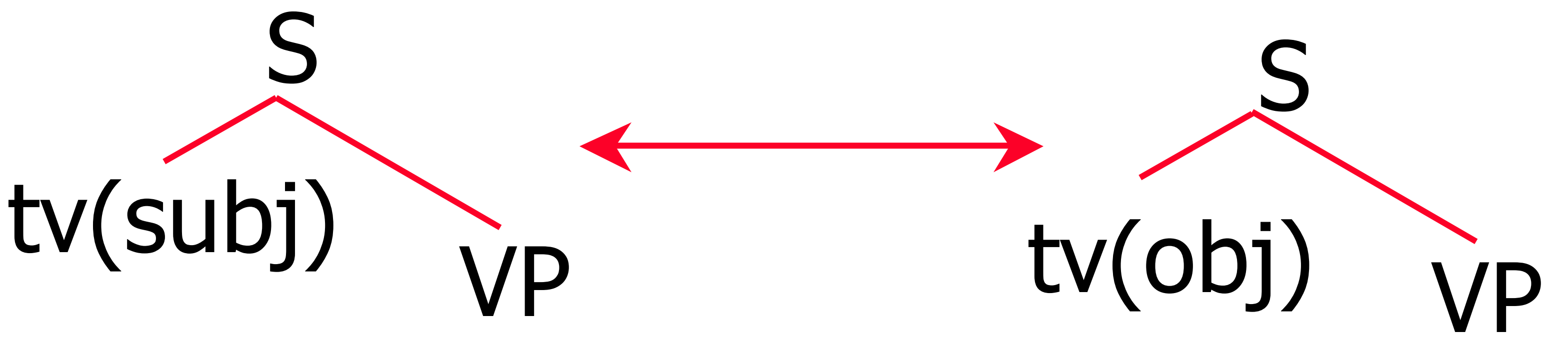


Handling other differences

- E: You like her
- S: Ella te gusta
- Lit: She you-accusative pleases

  (Grammatical object in English is subject in Spanish, and v.v.)

Tree mapping rule for this



Is this systematic?

- Yes, and taxonomic too…
- Roughly 8-9 such ‘classes’ of divergence:
  1. Thematic
  2. Head switching
  3. Structural
  4. Lexical Gap
  5. Lexicalization
  6. Categorial
  7. Collocational
  8. Multi-lexeme/idiomatic
  9. Generalization/morphological

Other divergences- systematic

- E: The baby just ate
- S: El bébé acaba de comer
- Lit: The baby finish of to-eat

  Head-switching

- E: Luisa entered the house
- S: Luisa entró a la casa
- Lit: Luisa entered to the house

  Structural

Divergences diverging

- E: Camilio got up early
- S: Camilio madrugó

  Lexical gap

- E: Susan swam across the channel
- S: Susan cruzó el canal nadando
- (Systran: Susan nadó a través del canal)
- Lit: Susan crossed the channel swimming

  (manner & motion combined in verb E, path in across; in S, verb cruzó has motion & path, motion in gerund nadnado)

  Lexicalization

Divergences, III

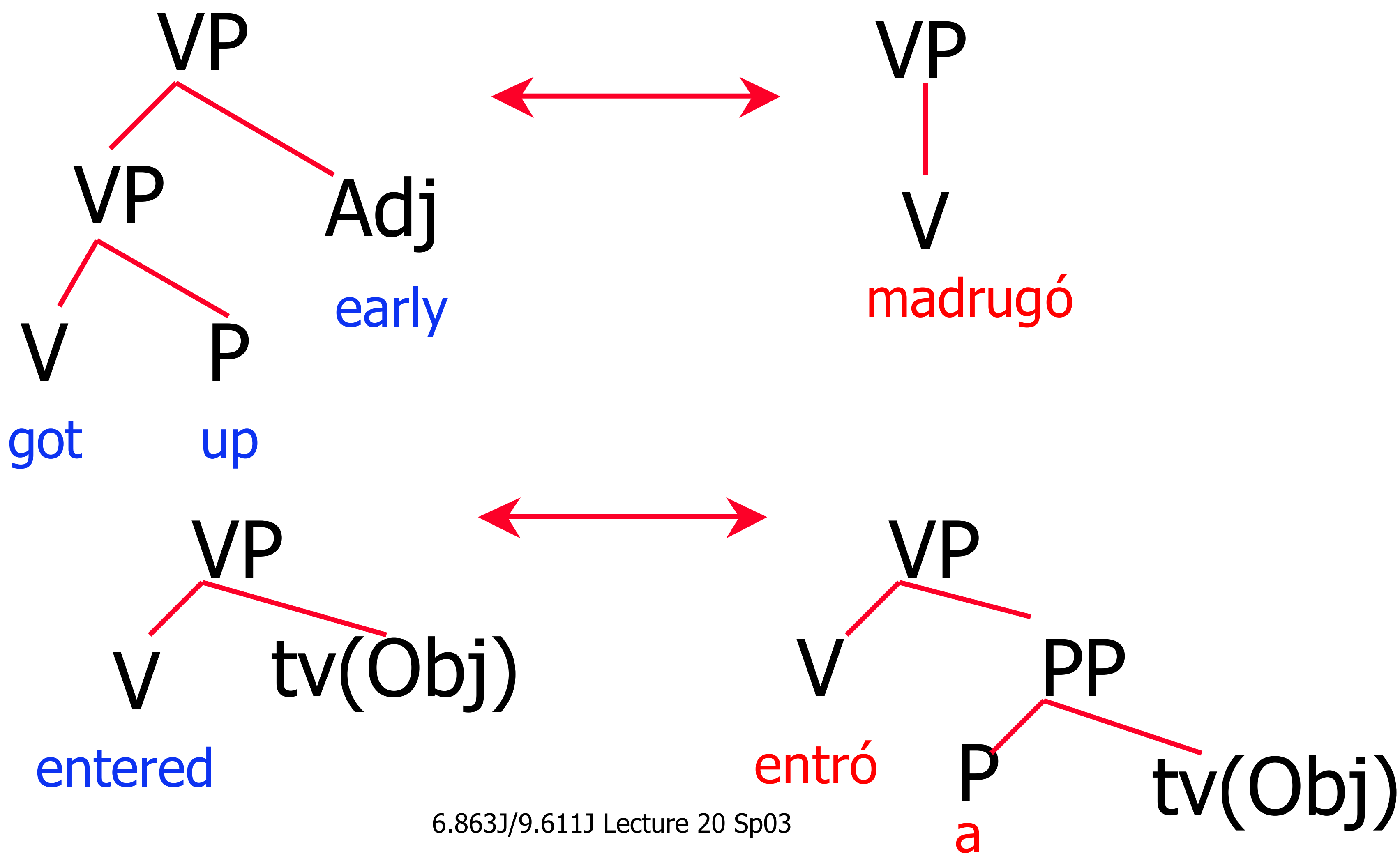
- E: A little bread
- S: Un poco de pan
- Lit: A bit of bread

Categorial – difft syntactic categories

- E: John made a decision
- S: John tomó/*hizo una decisión
- Lit: John took/*made a decision

Collocational – usually <u>make</u> goes to <u>hacer</u> but here a 'support' verb for decision

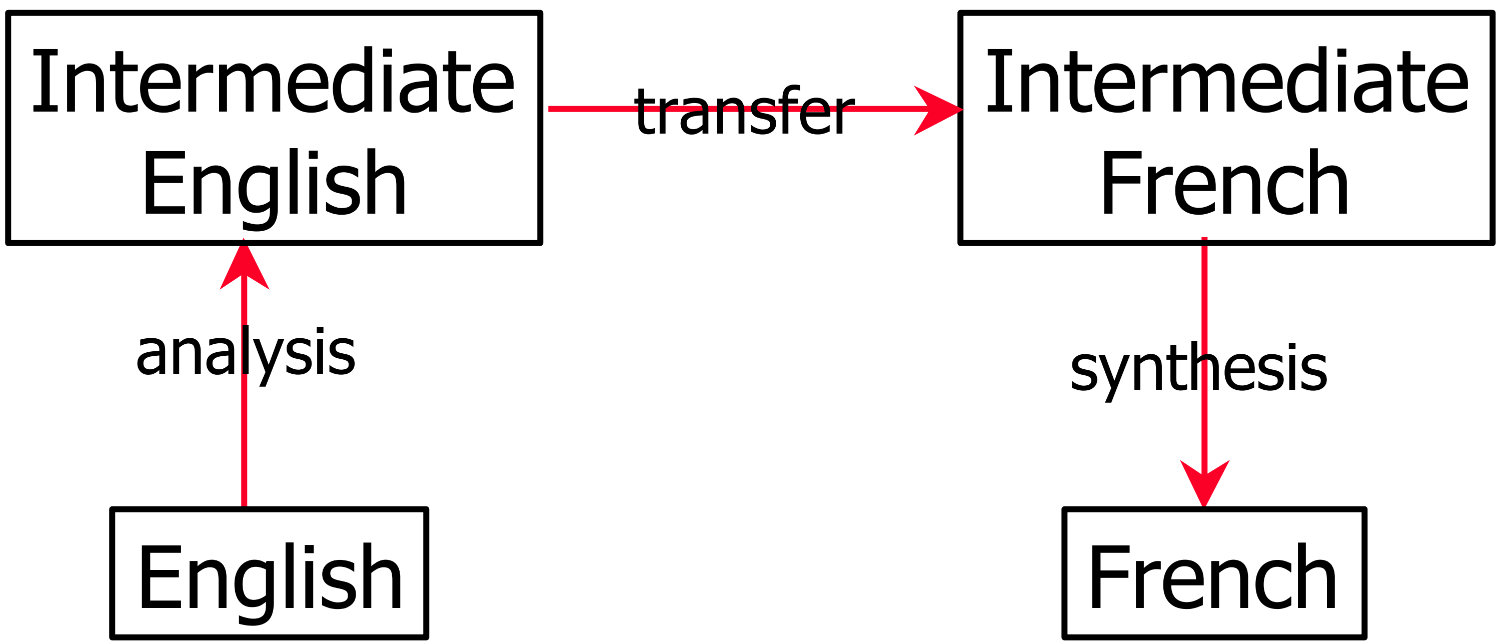
We can accommodate these…



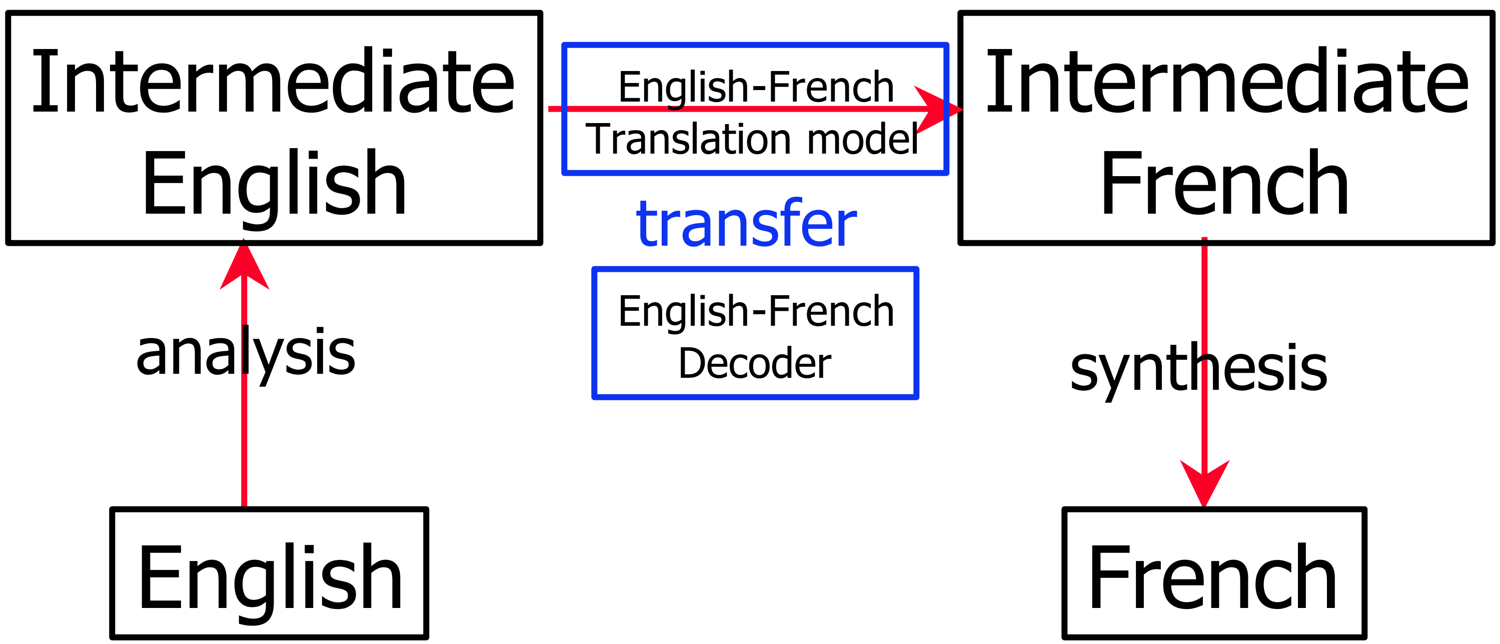
Issues

- Q: How many rules?
- A: usually many 000s for each one-way pair
- Q: Nondeterminism – which rule to apply?
- Q: How hard is it to build a rule system?
- A: Can we <u>learn</u> these automatically?

Transfer picture again



Statistical MT is transfer approach

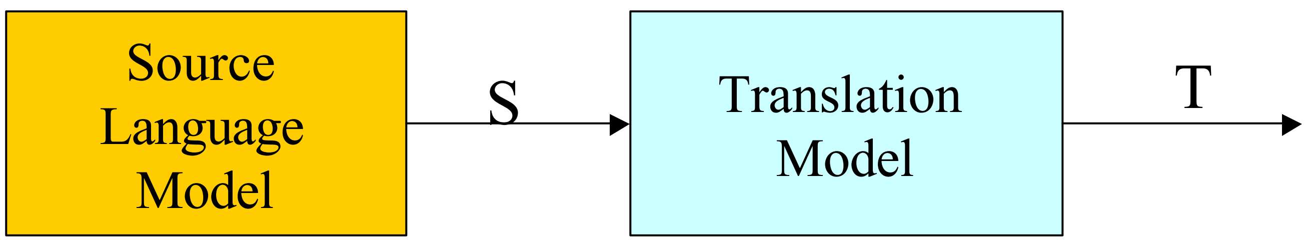


Statistical MT is transfer approach!

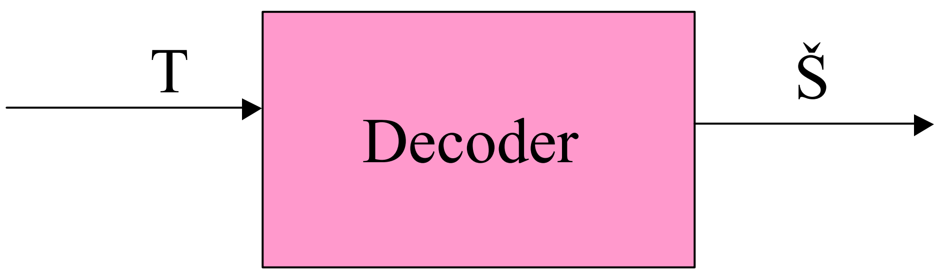
- Except…Analysis & synthesis vestigial
- Transfer done statistically
- Can we do better by applying some simple analysis & synthesis?
- A: Yes, from 50s to 60+ %
- A: Yes, we can do even better if we do incorporate syntax systematically: Knight et al 2001
- We will see that there's a method behind this madness, and all the alignment methods are in effect 'learning' the transfer rules

Adding syntax to Stat MT…simply

Statistical Machine Translation Model



$$P(S) \quad * \quad P(T|S) \quad = \quad P(S,T)$$



$$\check{S} = \operatorname*{argmax}_{S} P(S|T) = \operatorname*{argmax}_{S} P(S,T)$$

Brown et al, "A Statistical Approach to Machine Translation," 1990; 1993

Simple analysis and synthesis – IBM Model X

- Find word strings
- Annotate words via simple grammatical functions
- Very very very simple syntactic analysis
- Inflectional morphology
- Statistically derived word senses

Crummy but amazing improvement to stat model

- Simplest English & French syntactic regularization
- For English:
  - Undo question inversion
  - Move adverbs out of multiword verbs
  - Eg: Has the grocery store any eggs →

    The grocery store has any eggs Qinv →

    Iraq will probably not be completely balkanized→

    Iraq will be balkanized probably_m1 not_m2 completely_m3

And for French…

- Undo question inversion
- Combine *ne…pas*, *rien* into single words
- Move prounouns that function as direct, indirect, objects to position following verb & mark grammatical function
- Move adjs s.t. precede nouns they modify & adverbs to position following verbs they modify

French examples

- Où habite-il → Où il habite Qinv
- Il n'y en a plus → Il y en a ne_plus
- Je vous le donnerai → Je donnerai le_Pro vous_iPro ("I gave it to you")

How well does this work?

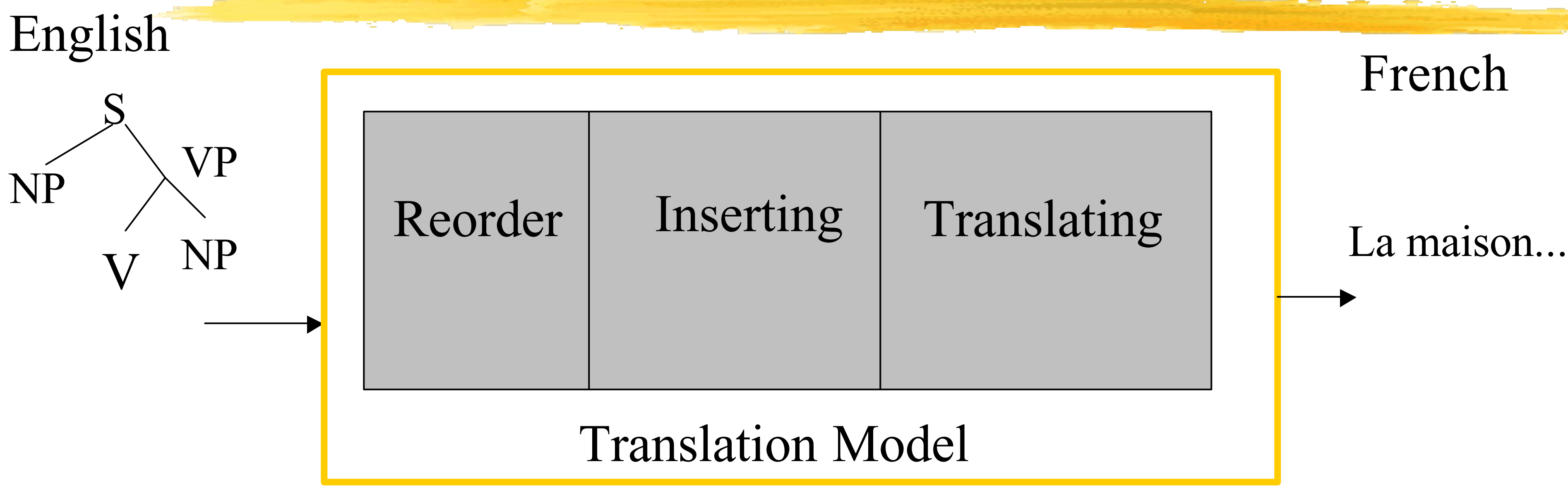
- Pretty darn well
- Improves performance about 10%
  - 50-odd % to 60+

- Now – let's see if we can reach the next step by doing this a bit more thoroughly:
- Add linguistic features to a statistical translation model by using parse trees

Add different 'channels' of noise

- Instead of *one* noisy channel, break it out into syntactic possibilities
- Reorder – model S V O vs. S OV order (Chinese, English vs. Japanese, Turkish)
- Insertion – model case particles
- Translating – as before

Syntax-based MT



Reorder	each node stochastically re-ordered N! possible re-orderings	*d-table*
Insert	syntactic case	*n-table*
Translation	word-by-word replacement	*t-table*

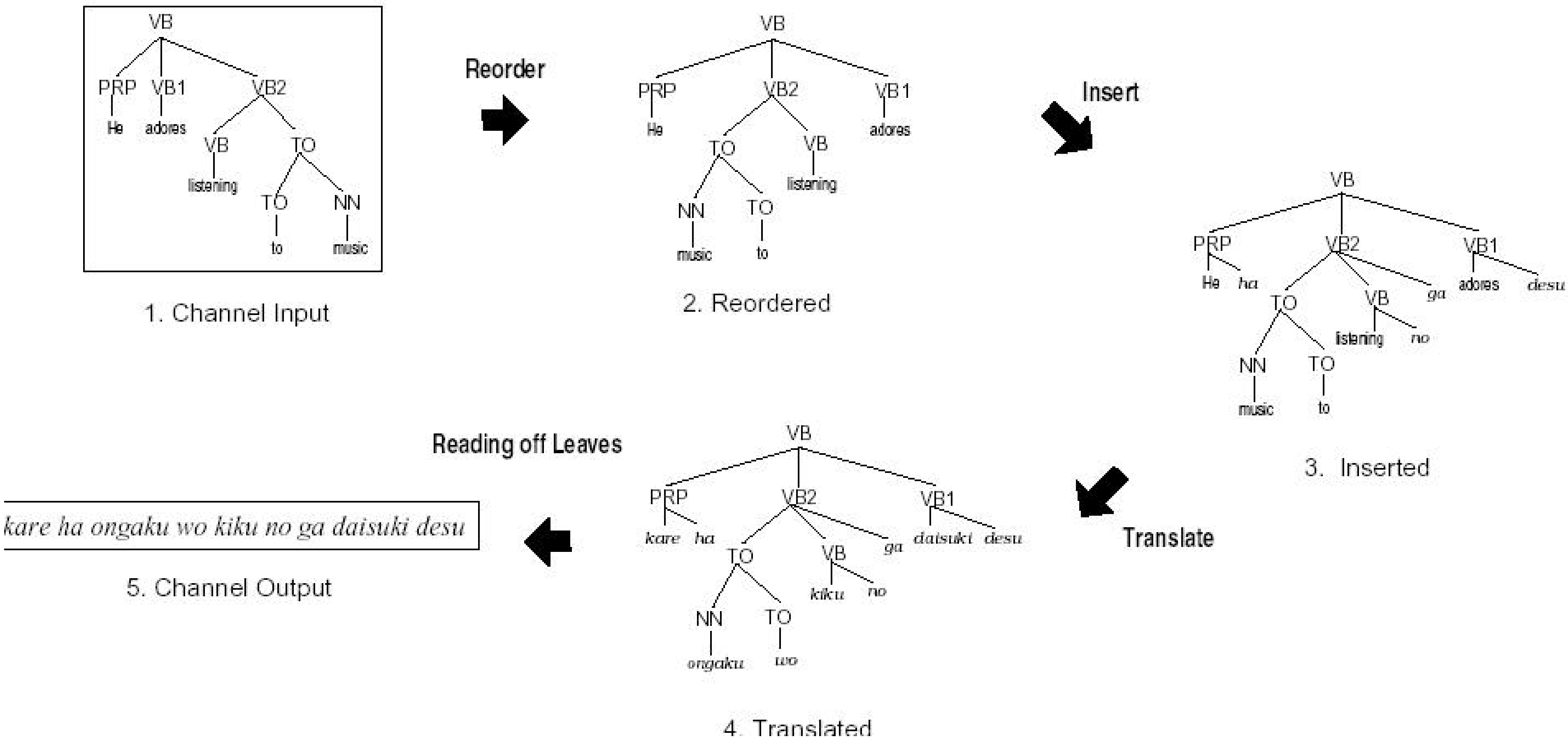
Sentence translation, E to J

He enjoys listening to music

↓

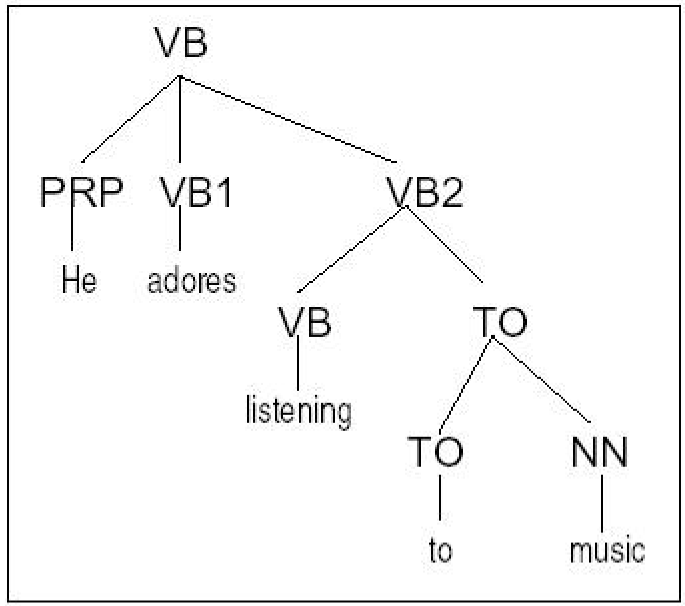
kare ha ongaku wo kiku no ga daisuki desu

Channeling



VB
PRP
VB1
VB2
He
adores
VB
TO
listening
TO
NN
to
music
1. Channel Input
Reorder
2. Reordered
Insert
He
ha
ga
adores
desu
listening
no
3. Inserted
Translate
kare
ha
daisuki
desu
kiku
no
ongaku
wo
4. Translated
Reading off Leaves
kare ha ongaku wo kiku no ga daisuki desu
5. Channel Output

Channeling - input



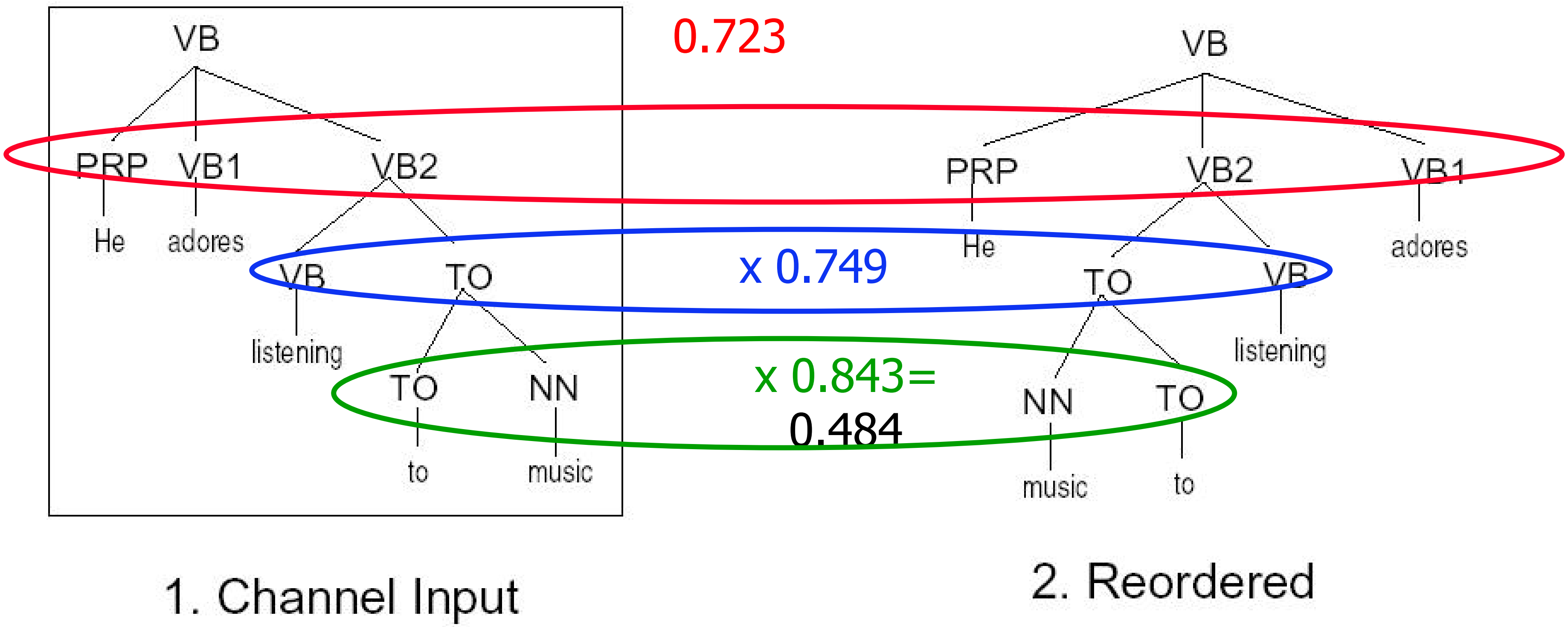
1. Channel Input

Reordering (r-table)

original order	reordering	P(reorder)
PRP VB1 VB2	PRP VB1 VB2	0.074
	PRP VB2 VB1	0.723
	VB1 PRP VB2	0.061
	VB1 VB2 PRP	0.037
	VB2 PRP VB1	0.083
	VB2 VB1 PRP	0.021
VB TO	VB TO	0.251
	TO VB	0.749
TO NN	TO NN	0.107
	NN TO	0.893
⋮	⋮	⋮

r−table

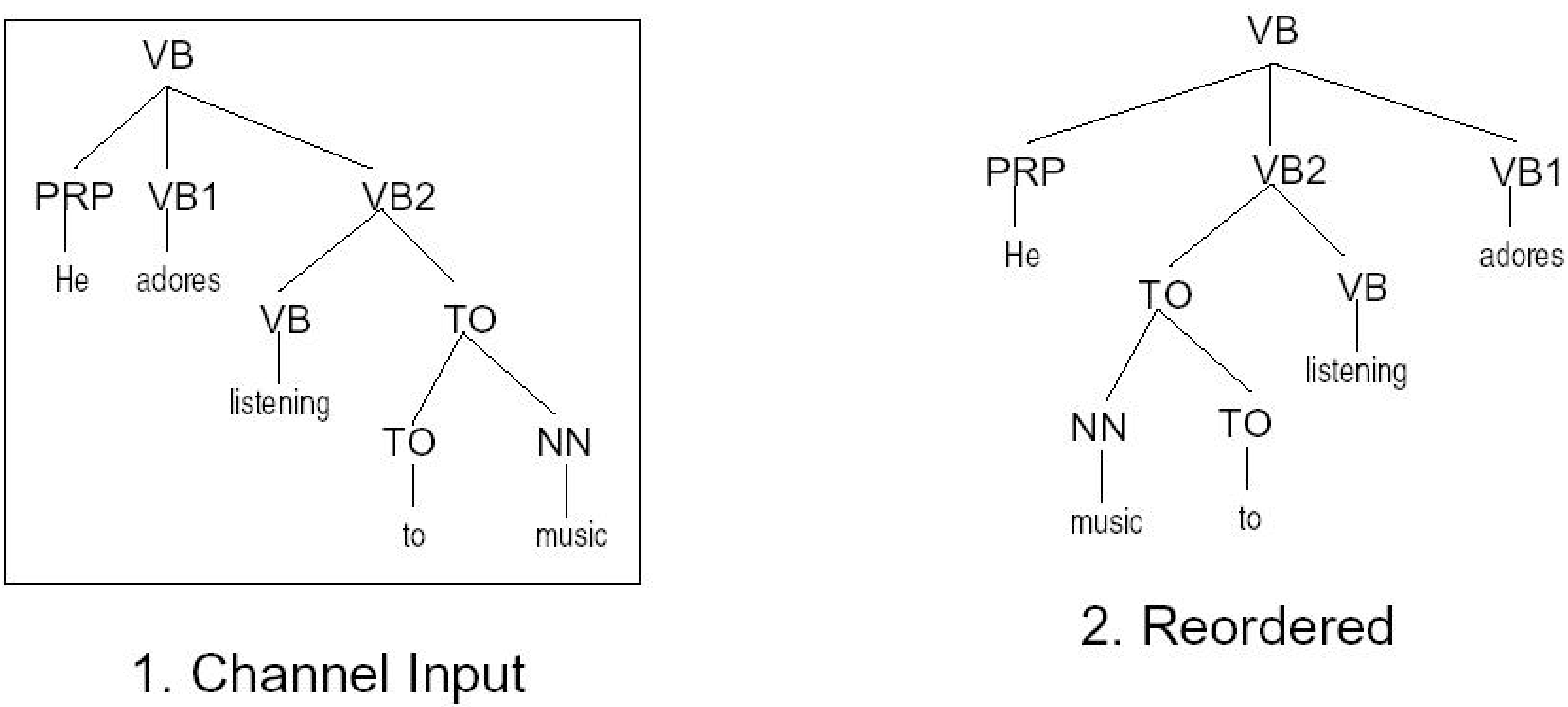
Reordered


VB
PRP
VB1
VB2
He
adores
VB
TO
listening
TO
NN
to
music
0.723
x 0.749
x 0.843=
0.484
VB
PRP
VB2
VB1
He
adores
TO
VB
listening
NN
TO
music
to
1. Channel Input
2. Reordered

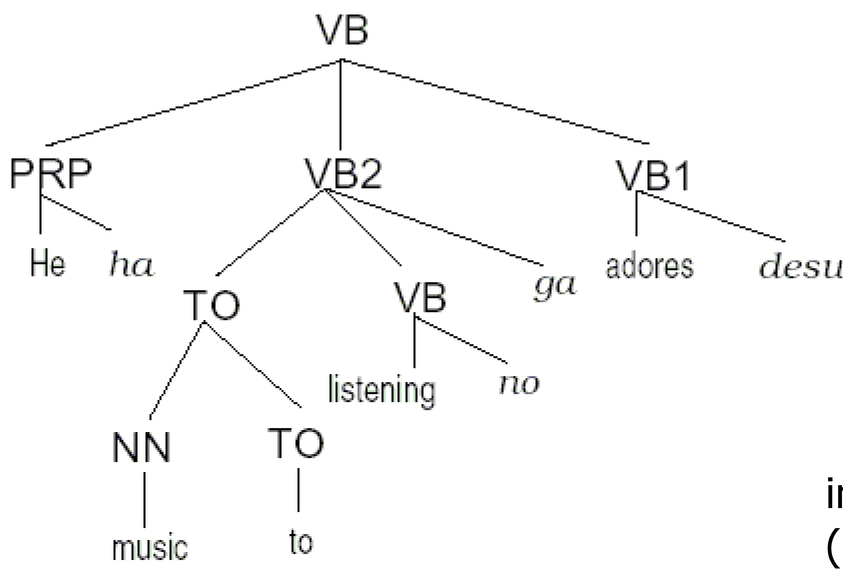
Channeling

- child nodes on each internal node are reordered, via R-table
- Eg: PRP-VB1-VB2 to PRP-VB2-VB1 has pr 0.723, so we pick that one
- Also reorder VB-TO $\rightarrow$ TO-VB; TO-NN $\rightarrow$NN-TO
- Prob of the 2nd tree is therefore 0.723 x 0.749 x 0.893 = 0.484

Reordered


VB
PRP
VB1
VB2
He
adores
VB
TO
listening
TO
NN
to
music
1. Channel Input
VB
PRP
VB2
VB1
He
TO
VB
adores
listening
NN
TO
music
to
2. Reordered

Insertion



3. Inserted

inserted four words
(*ha, no, ga* and *desu*)
to create the third tree
The top VB node, two TO nodes,
and the NN node inserted nothing

Insertion

- Captures regularity of inserting case markers ga, wa, etc.
- No conditioning – case marker just as likely anyplace

Insertion – n-table

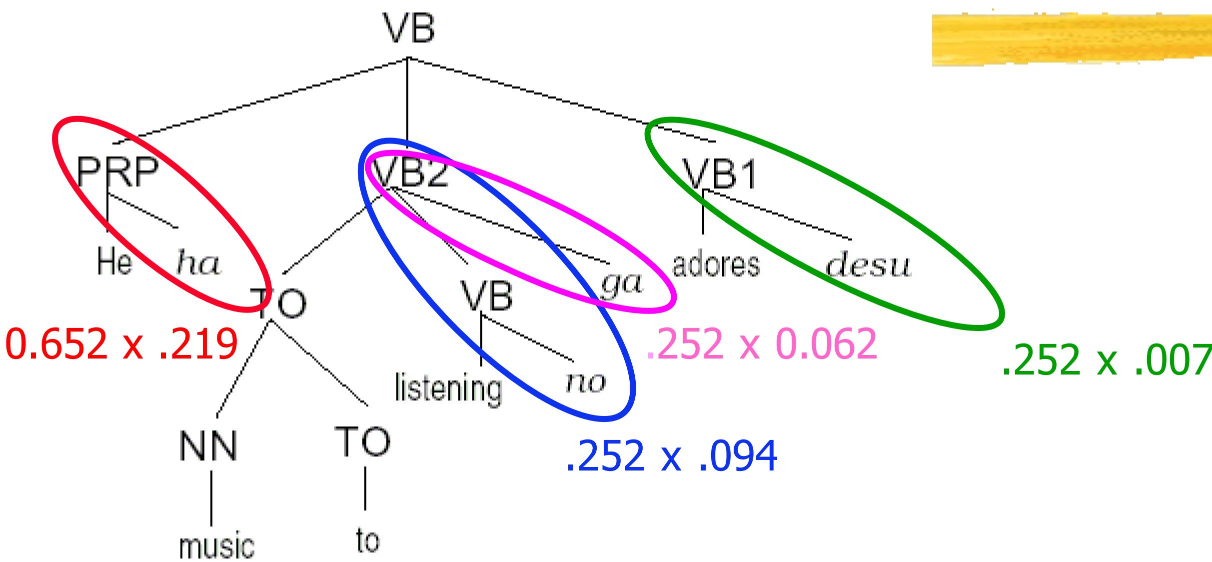
- Left, right, or nowhere (difft from IBM)
- 2-way table index, by (node, parent)
- EG, PRP node has parent VB

parent	TOP	VB	VB	VB	TO	TO	...
node	VB	VB	PRP	TO	TO	NN	...
P(None)	0.735	0.687	0.344	0.709	0.900	0.800	...
P(Left)	0.004	0.061	0.004	0.030	0.003	0.096	...
P(Right)	0.260	0.252	0.652	0.261	0.007	0.104	...

Insertion – which words to insert table

w	P(ins–w)
ha	0.219
ta	0.131
wo	0.099
no	0.094
ni	0.080
te	0.078
ga	0.062
⋮	⋮
desu	0.0007
⋮	⋮

Insertion



3. Inserted

inserted four words
(*ha, no, ga* and *desu*)
to create the third tree
The top VB node, two TO nodes,
and the NN node inserted nothing

So, probability of obtaining the third tree
given the second tree is: 4 particles x no inserts =

ha	no	ga	desu

(0.652 x .219)(0.252 x 0.094)(0.252 x 0.062)(0.252 x 0.007)x
0.735 x 0.709 x 0.900 x 0.800 = 3.498e-9

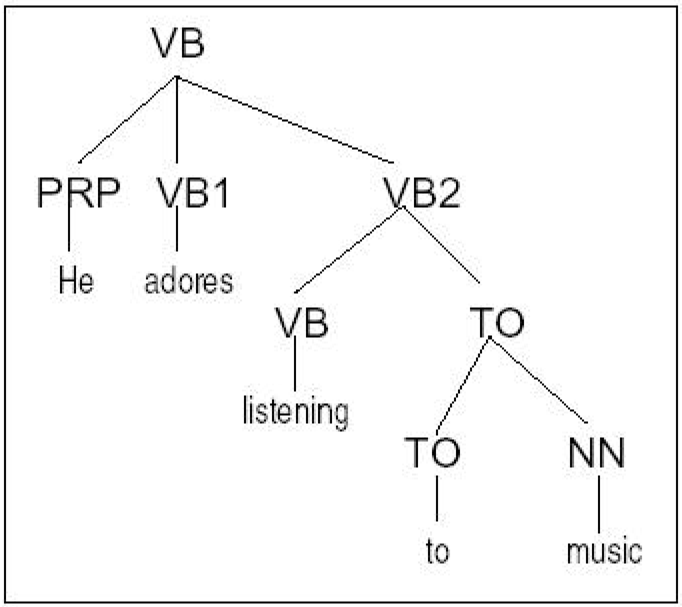
Translate – final channeling

- Apply the *translate* operation to each leaf
- Dependent only on the word itself and that no context
- Translations for the tree shown…

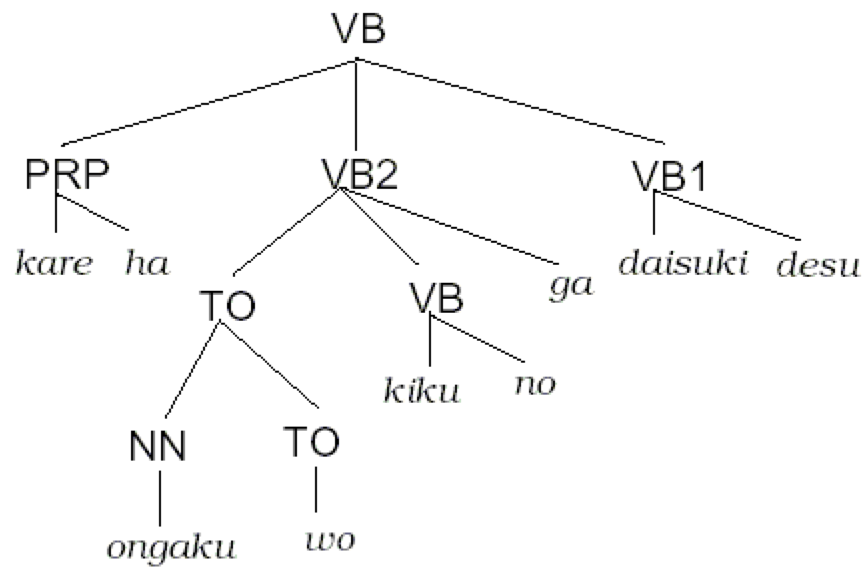
Translation, t-table

E	adores		he		i		listening		music		to		···
J	*daisuki*	1.000	*kare*	0.952	*NULL*	0.471	*kiku*	0.333	*ongaku*	0.900	*ni*	0.216	···
			NULL	0.016	*watasi*	0.111	*kii*	0.333	*naru*	0.100	*NULL*	0.204	
			nani	0.005	*kare*	0.055	*mi*	0.333			*to*	0.133	
			da	0.003	*shi*	0.021					*no*	0.046	
			shi	0.003	*nani*	0.020					*wo*	0.038	
			⋮	⋮	⋮	⋮					⋮	⋮	

Translated tree

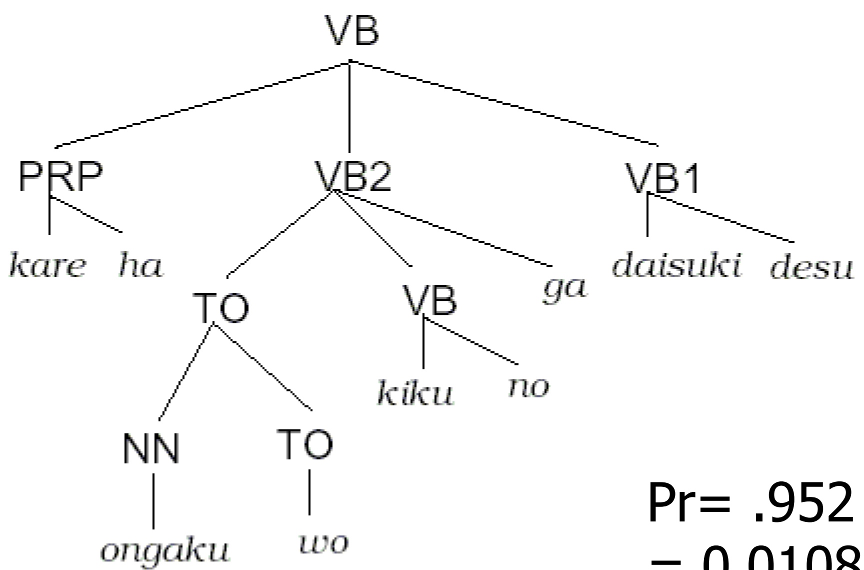


1. Channel Input



4. Translated

Translated tree



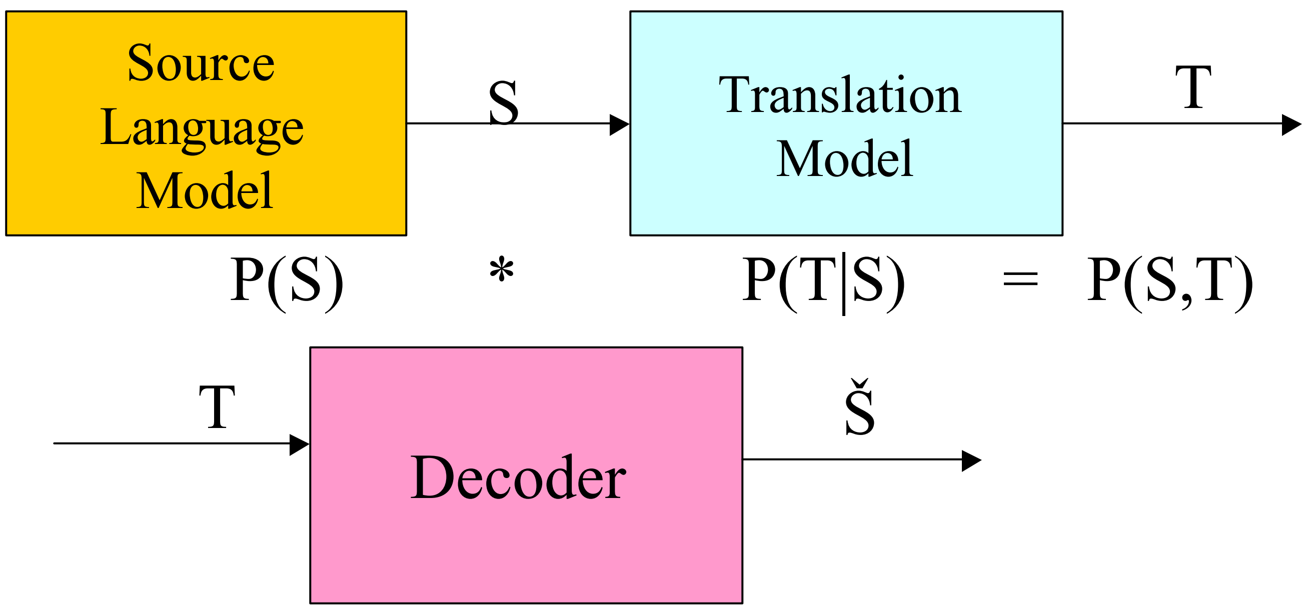
Pr= .952 x .900 x .038 x 1
= 0.0108

4. Translated

Total probability for this (e,j) pair

- Product of these 3 ops
- But many other combinations of these 3 ops yield same japanese sentence, so must sum these pr's…
- Actually done with 2121 E/J sentence pairs
- Uses efficient implementation of EM (50 mins per iteration, 20 iterations)

Statistical Machine Translation Model



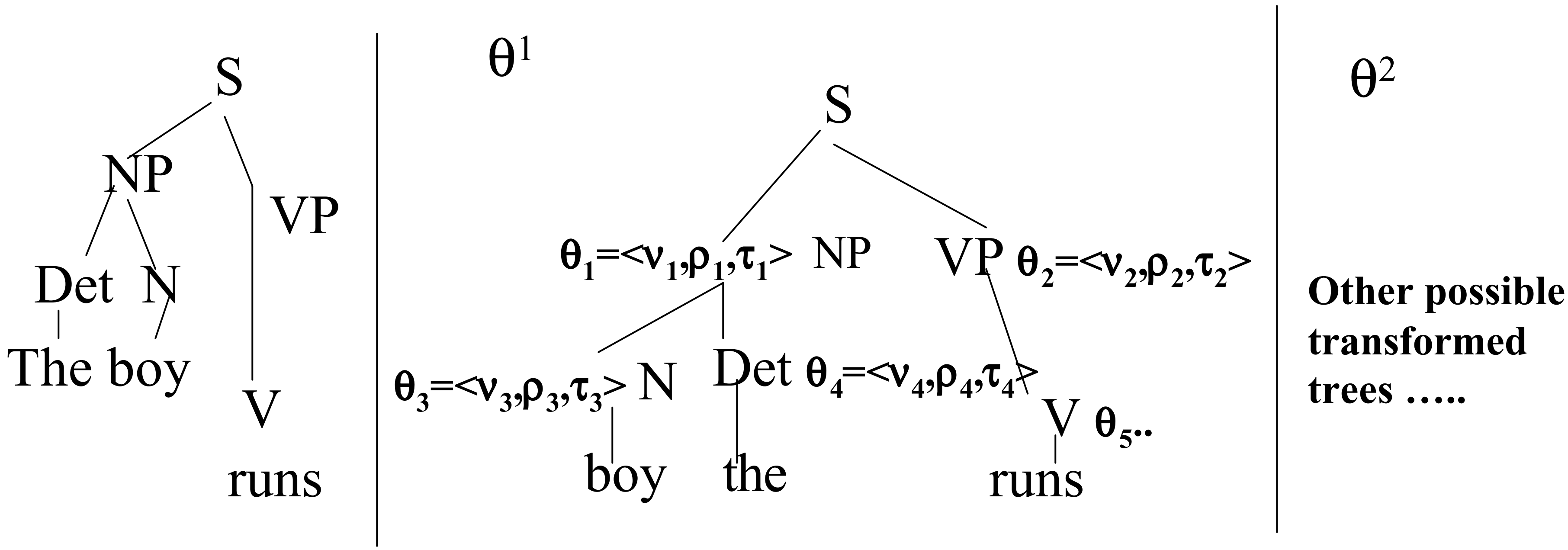
$$\check{S} = \underset{S}{\operatorname{argmax}}\, P(S|T) = \underset{S}{\operatorname{argmax}}\, P(S,T)$$

Brown et al, "A Statistical Approach to Machine Translation," 1990; 1993

Syntax-based MT

Random variables N, R, T each representing one of the channel operations for each (E)nglish node ε

Insertion N (ν) - *Reorder* R (ρ) - *Translation* T (τ)



Parameters of this model

$$P(f \mid \varepsilon) = \sum_{\theta : Str(\theta(\varepsilon)) = f} P(\theta \mid \varepsilon)$$

$$P(\theta \mid \varepsilon) = \prod_{i=1}^{n} P(\theta_i \mid \varepsilon_i)$$

$$P(\theta_i \mid \varepsilon_i) = P(\nu_i, \rho_i, \tau_i \mid \varepsilon_i)$$

$$= P(\nu_i \mid \varepsilon_i) P(\rho_i \mid \varepsilon_i) P(\tau_i \mid \varepsilon_i)$$

$$= P(\nu_i \mid \mathrm{N}(\varepsilon)_i) P(\rho_i \mid R(\varepsilon)_i) P(\tau_i \mid T(\varepsilon_i))$$

$$= n(\nu_i \mid \mathrm{N}(\varepsilon_i)) r(\rho_i \mid R(\varepsilon_i)) t(\tau_i \mid T(\varepsilon_i))$$

Now do EM magic

$$P(f \mid \varepsilon) = \sum_{\theta : Str(\theta(\varepsilon)) = f} \prod_{i=1}^{n} n(\nu_i \mid \mathrm{N}(\varepsilon)_i) r(\rho_i \mid R(\varepsilon)_i) t(\tau_i \mid T(\varepsilon_i))$$

EM

- initialize model parameters
- Repeat
  - E probabilities of the events are calculated from current model parameters
  - M number of events are weighted with the probabilities of the events
- re-estimate model parameters based on observed counts

Parameter estimation via EM

EM:

1. Initialize all probability tables: $n(\upsilon, N)$ $r(\rho, R)$ and $t(\tau,T)$
2. Reset all counters $c(\upsilon, N)$ $c(\rho, R)$ and $c(\tau,T)$

E

3. For each pair $<\varepsilon, f>$ in the training corpus
   - For all θ, such that $f = \text{String}(\theta(\varepsilon))$,
     - •Let $cnt = P(\theta|\varepsilon)/\Sigma_{\theta:\text{Str}(\theta(\varepsilon))=f,} P(\theta|\varepsilon)$
     - •For $i = 1 \ldots n$
       - $c(\upsilon_i, N(\varepsilon_i))$ += cnt
       - $c(\rho_i, R(\varepsilon_i))$ += cnt
       - $c(\tau_i, T(\varepsilon_i))$ += cnt

M

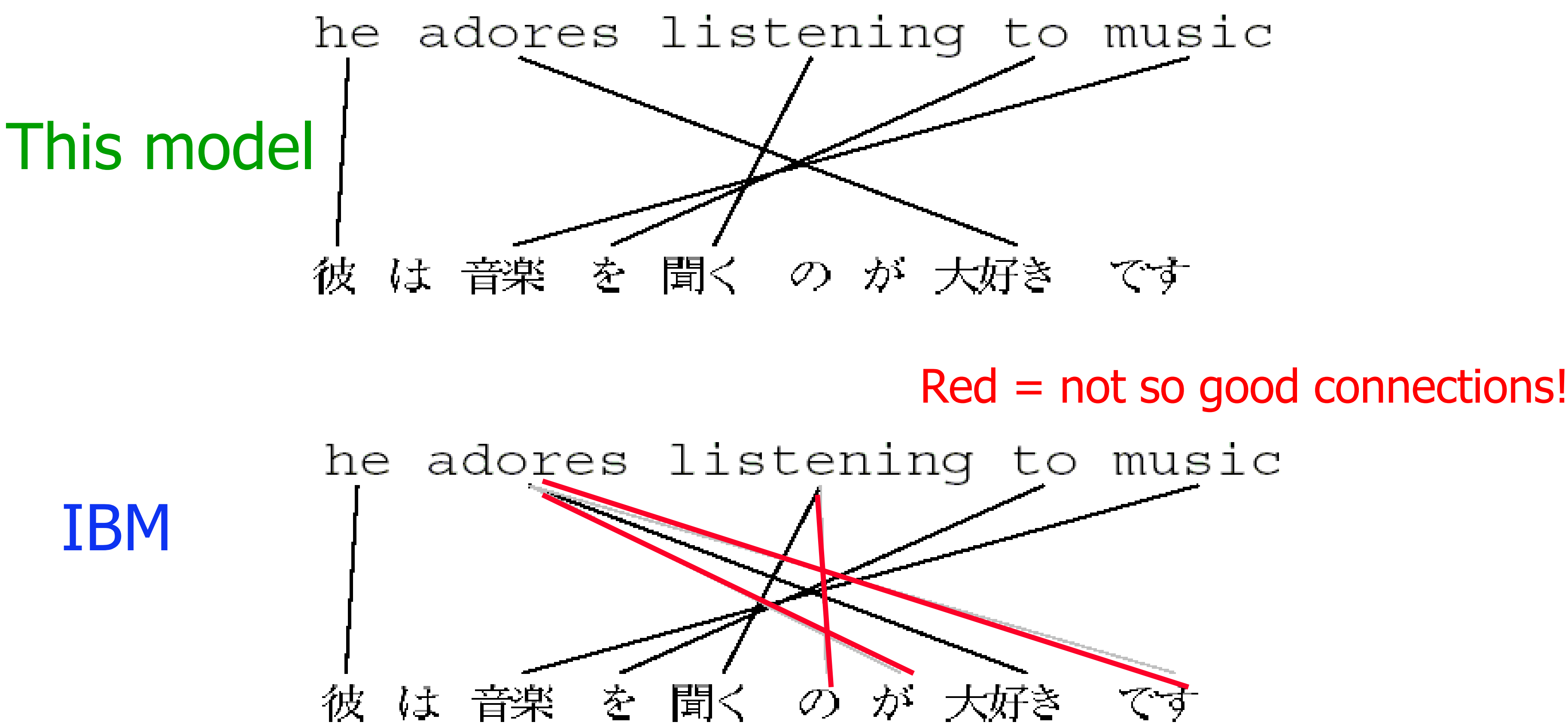
4. For each (υ, N) (ρ, R) and (τ,T),
   - $n(\upsilon, N) = c(\upsilon, N) / \Sigma_v c(\upsilon, N)$
   - $r(\rho, R) = c(\rho, R) / \Sigma_p c(\rho, R)$
   - $t(\tau,T) = c(\tau,T) / \Sigma_t c(\tau,T)$
5. Repeat steps 2-4 for several iterations (until little change) [20 steps]

Parameter estimation via EM

$\mathbf{O}(\ |\upsilon|^{n}\ |\rho|^{n}\ )$
for all possible combinations of parameters (υ,ρ,τ)

$\mathbf{O}\ (\ n^{3|}\ |\ \upsilon\ |\ |\rho|\ |\pi|\ )$

Results vs. IBM Model 5



Results for 50 sentence pairs

Perfect = all alignments OK for 3 judges
Scoring: 1.0 = OK; 0.5 = not sure; 0= wrong

	Alignment ave. score	Perfect sents
Our Model	0.582	10
IBM Model 5	0.431	0

For E-F, goes up to 70%!

Can we get to the next step up – "Gold Standard" of 80%??

Problemos

- F in: L’atmosphère de la Terre rend un peu myopes mêmes les meilleurs de leur télèscopes
- E out: The atmosphere of the Earth returns a little myopes same the best ones of their telescopes
- (Systran): The atmosphere of the Earth makes a little short-sighted same the best of their télèscopes
- (Better) The earth’s atmosphere makes even the best of their telescopes a little ‘near sighted’
- Why?

Pourquois?

- French verb <u>rend</u> can be ‘return’ or ‘make’
- French word même can be ‘same’ or ‘even’ – translation systems get it dead wrong

Problem of context

- General vs. specialized use of word
- "Dear Bill," to German:
- Liebe Rechnung –
- "beloved invoice"
- (Systran) Liebe Bill
- Solution: consult word senses?

Anaphora and beyond…

- Die Europäische Gemeinschaft und ihre Mitglieder
  - The European Community and its members
  - Die Europäische Gemeinschaft und seine Mitglieder
- The monkey ate the banana because it was hungry
  - Der Affe ass die Banane weil er Hunger hat
  - Der Affe aß die Banane, weil sie hungrig war
- The monkey ate the banana because it was ripe
  - Der Affe ass die Banane weil sie reif war
- The monkey ate the banana because it was lunch-time
  - Der Affe ass die Banane weil es Mittagessen war
- Sentence-orientation of all systems makes most anaphora problematic (unresolvable?); possibly a discourse-oriented ‘language model’ is the only chance