Source-side Dependency Tree Reordering Models with Subtree Movements and Constraints

Nguyen Bach, Qin Gao and Stephan Vogel

Carnegie Mellon University

interACT

International Center for Advanced Communication Technologies

Carnegie Mellon

lti

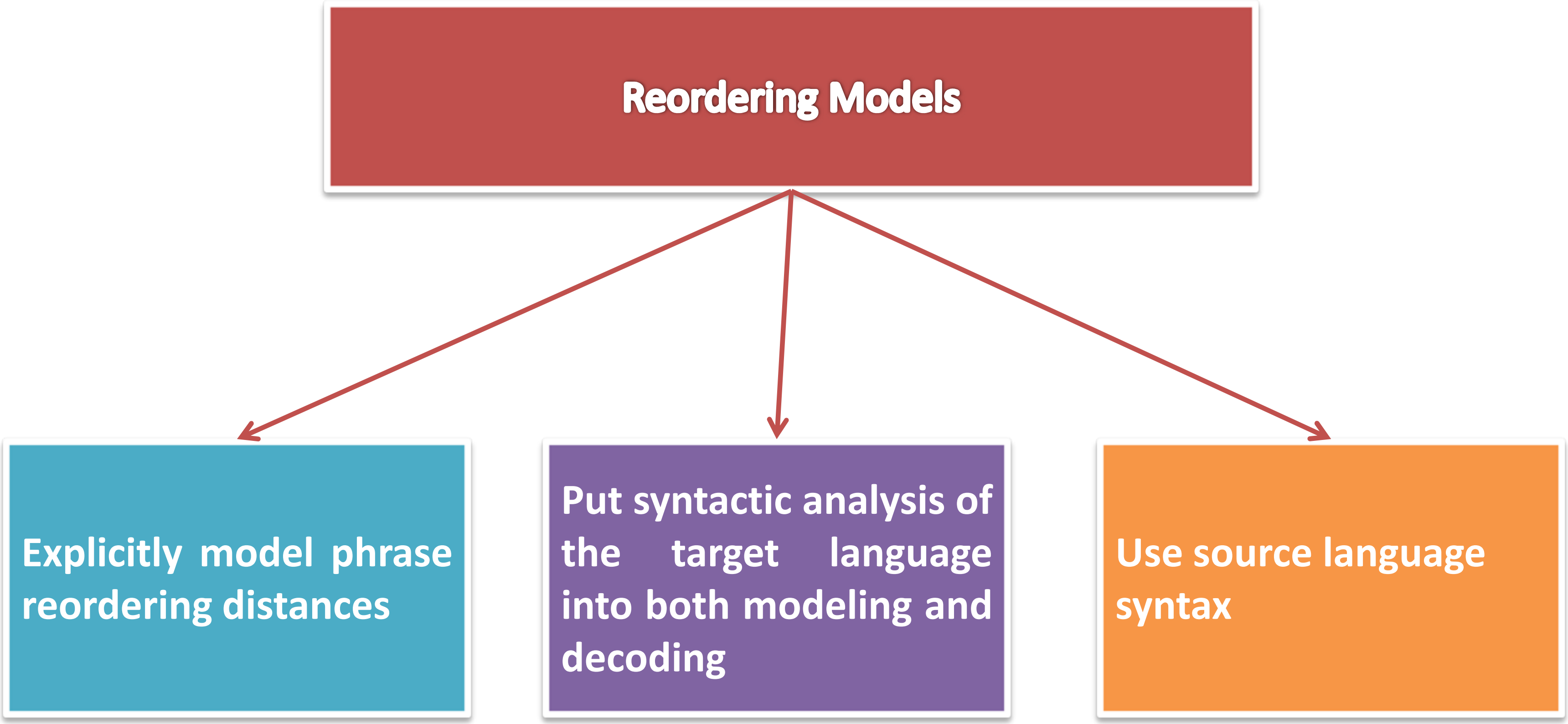
Overview

- We introduce **source-side dependency tree reordering models**
  - Inspired by **lexicalized reordering model** (Koehn et. al 2005) , **hierarchical dependency translation** (Shen et. al, 2008) and **cohesive decoding** (Cherry, 2008)
- We model reordering events of phrases associated with source-side dependency trees
  - **Inside/Outside subtree movements** efficiently capture the statistical distribution of the subtree-to-subtree transitions in training data
  - Utilize subtree movements directly at the decoding time alongside with **cohesive constraints** to guide the search process
  - Improvements are shown in English-Spanish and English-Iraqi tasks

Outline

- **Background & Motivations**
- Source-side dependency tree reordering models
  – Modeling
  – Training
  – Decoding
- Experiments & Analysis
- Conclusions

Background of Reordering Models



Reordering Models

- **Explicitly model phrase reordering distances**
  - Distance-based (Och, 2002; Koehn et.al., 2003)
  - Lexicalized phrase (Tillmann, 2004; Koehn, et.al., 2005; Al-Onaizan and Papineni, 2006)
  - Hierarchical phrase (Galley and Manning, 2008)
  - MaxEnt classifier (Zens and Ney, 2006; Xiong, et.al., 2006; Chang, et. al., 2009)
- **Put syntactic analysis of the target language into both modeling and decoding**
  - Direct model target language constituents movement in either constituency trees (Yamada and Knight, 2001; Galley et.al., 2006; Zollmann et.al., 2008) or dependency trees (Quirk, et.al., 2005)
  - Hierarchical phrase-based (Chiang, 2005; Shen et. al., 2008)
- **Use source language syntax**
  - Preprocessing with syntactic reordering rules (Xia and McCord, 2004; Collins et.al., 2005; Rottmann and Vogel, 2007; Wang et.al., 2007; Xu et.al. 2009)
  - Use syntactical analysis to provide multiple source sentence reordering options through word lattices (Zhang et.al., 2007; Li et.al., 2007; Elming, 2008).

Reordering Models

Explicitly model phrase reordering distances

Put syntactic analysis of the target language into both modeling and decoding

Use source language syntax

Distar
Koehr

Lexica
2004;
Onaiz

Hierar
and M

MaxEnt classifier (Zens and Ney, 2006; Xiong, et.al., 2006; Chang, et. al., 2009)

Hierarchical phrase-based (Chiang, 2005; Shen et. al., 2008)

ctic
al.,
el,
Xu

o
ons
ang
et.al., 2007; Li et.al., 2007; Elming, 2008).

Source-side **Dependency Tree** **Reordering Models**

with Subtree Movements and Constraints

What are the differences?

- Instead of using flat word structures to extract reordering events, utilize **source-side dependency structures**
  - Provide more linguistic cues for reordering events
- Instead of using pre-defined reordering patterns, learn **reordering feature distributions** from training data
  - Capture reordering events from real data
- Instead of preprocessing the data, **discriminatively train** the reordering model via MERT
  - Tighter integration with the decoder

Cohesive Decoding

- A cohesive decoding (Cherry, 08; Bach et. al., 09) is forcing the cohesive constraint:
  - *When the decoder begins translation any part of a source subtree, it must cover all words under that subtree before it can translate anything outside.*

- Source-side dependency tree reordering models
  - Efficiently capture the statistical distribution of the subtree-to-subtree transitions in training data.
  - Directly utilize it at the decoding time to guide the search process.

Outline

- Background of Reordering Models
- **Source-side dependency tree reordering models**
  - **Modeling**
  - **Training**
  - **Decoding**
- Experiments & Analysis
- Conclusions

Lexicalized Reordering Models (Tillmann, 2004; Koehn, et.al., 2005; Al-Onaizan & Papineni, 2006)

$$p(O \mid e, f) = \prod_{i=1}^{n} p(o_i \mid \bar{e}_i, \bar{f}_{a_i}, a_{i-1}, a_i)$$

where

f is the input sentence;

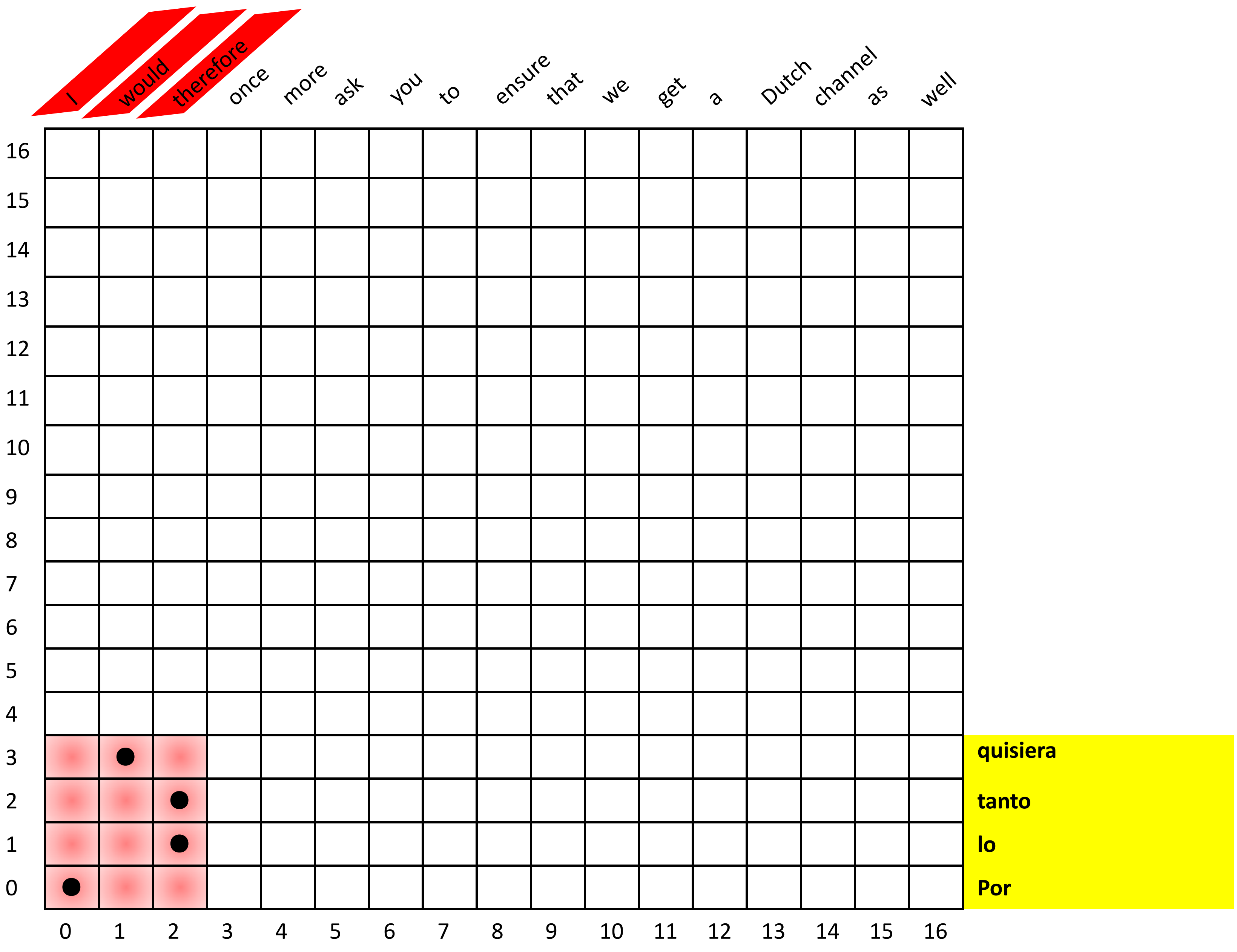
$e = (\bar{e}_1, \ldots, \bar{e}_n)$ is the target language phrases;

$a = (a_1, \ldots, a_n)$ is phrase alignments;

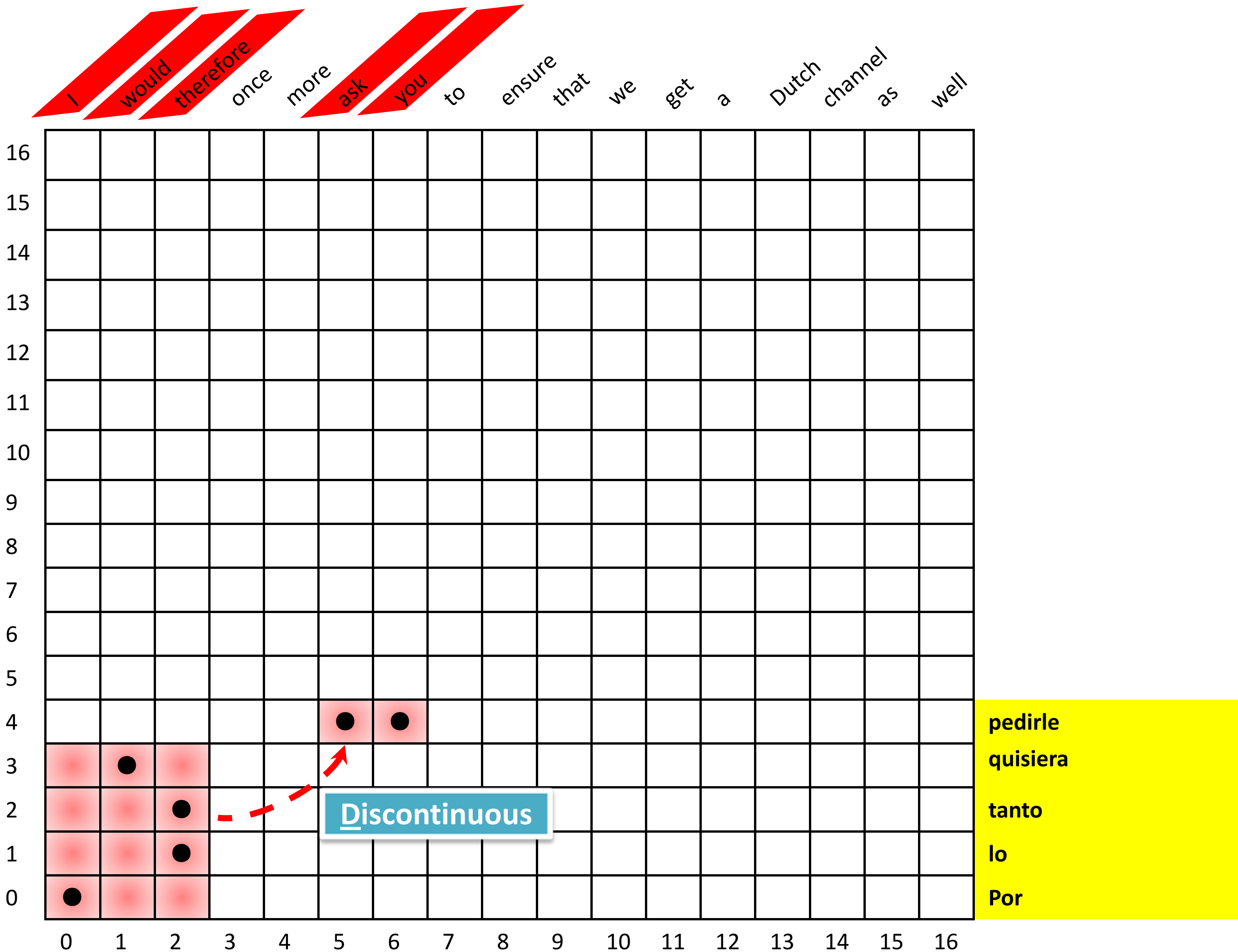
$\bar{f}_{a_i}$ is a source phrase which has a translated phrase $\bar{e}_i$ defined by an alignment a_i;

O is orientation phrase sequence; each o_i has a value over 3 possibles (M, S,D);

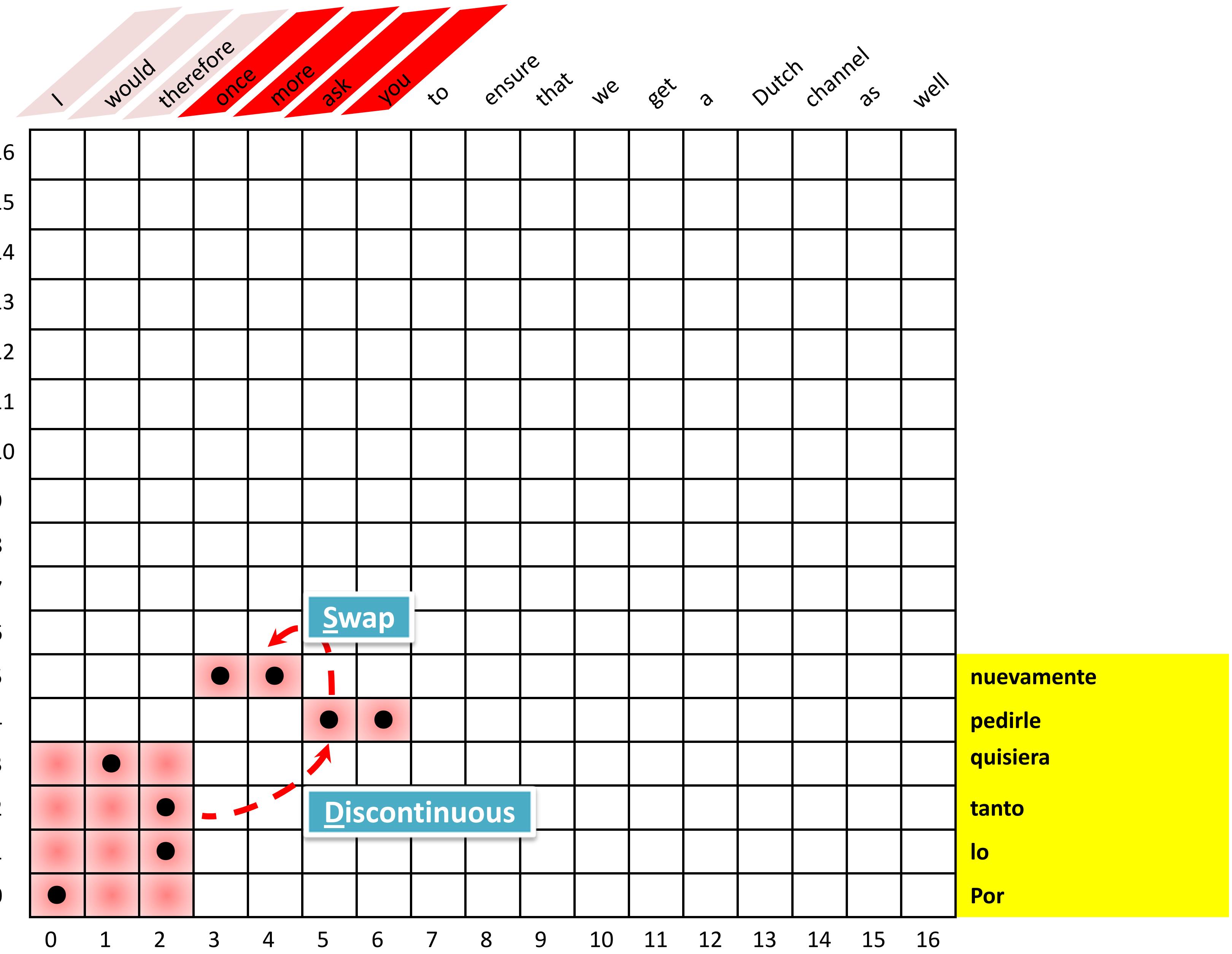
	I	would	therefore	once	more	ask	you	to	ensure	that	we	get	a	Dutch	channel	as	well
16																	
15																	
14																	
13																	
12																	
11																	
10																	
9																	
8																	
7																	
6																	
5																	
4																	
3																	
2																	
1																	
0																	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16



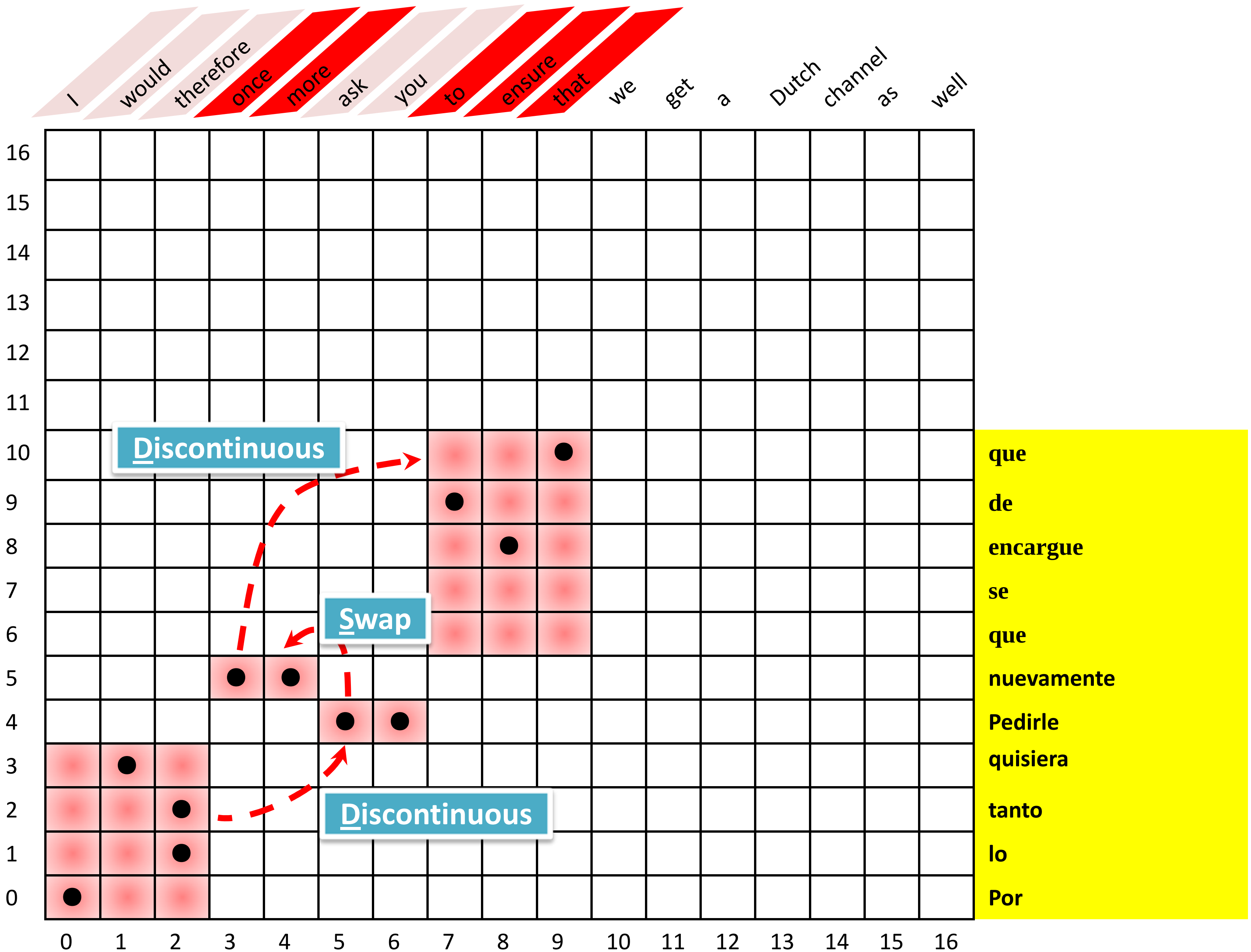
I
would
therefore
once
more
ask
you
to
ensure
that
we
get
a
Dutch
channel
as
well
quisiera
tanto
lo
Por
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

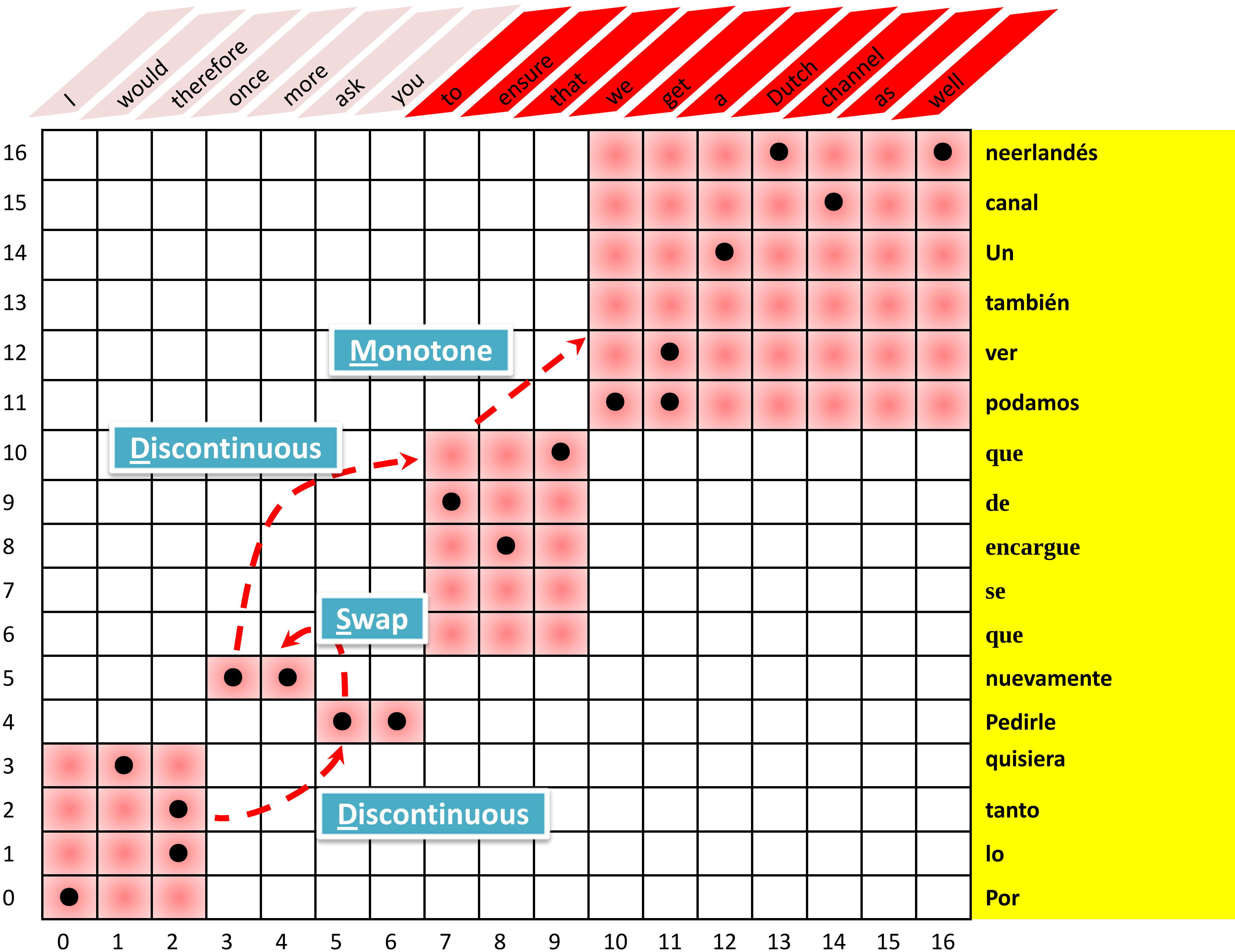


I
would
therefore
once
more
ask
you
to
ensure
that
we
get
a
Dutch
channel
as
well
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Discontinuous
pedirle
quisiera
tanto
lo
Por



I
would
therefore
once
more
ask
you
to
ensure
that
we
get
a
Dutch
channel
as
well
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
Swap
Discontinuous
nuevamente
pedirle
quisiera
tanto
lo
Por
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16


I would therefore once more ask you to ensure that we get a Dutch channel as well
que
de
encargue
se
que
nuevamente
Pedirle
quisiera
tanto
lo
Por
Discontinuous
Swap
Discontinuous
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16

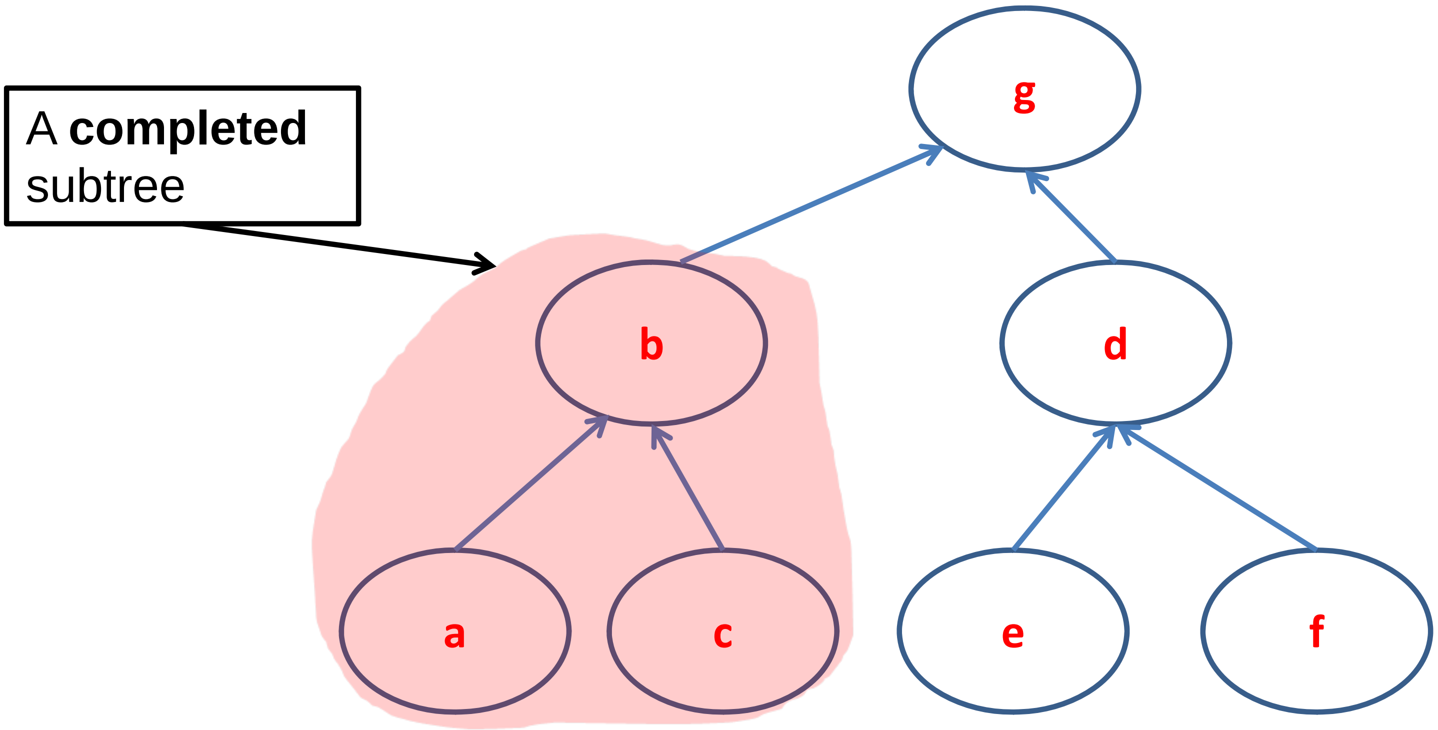


I
would
therefore
once
more
ask
you
to
ensure
that
we
get
a
Dutch
channel
as
well
Monotone
Discontinuous
Swap
Discontinuous
neerlandés
canal
Un
también
ver
podamos
que
de
encargue
se
que
nuevamente
Pedirle
quisiera
tanto
lo
Por

Pros & Cons of Lexicalized Reordering Models

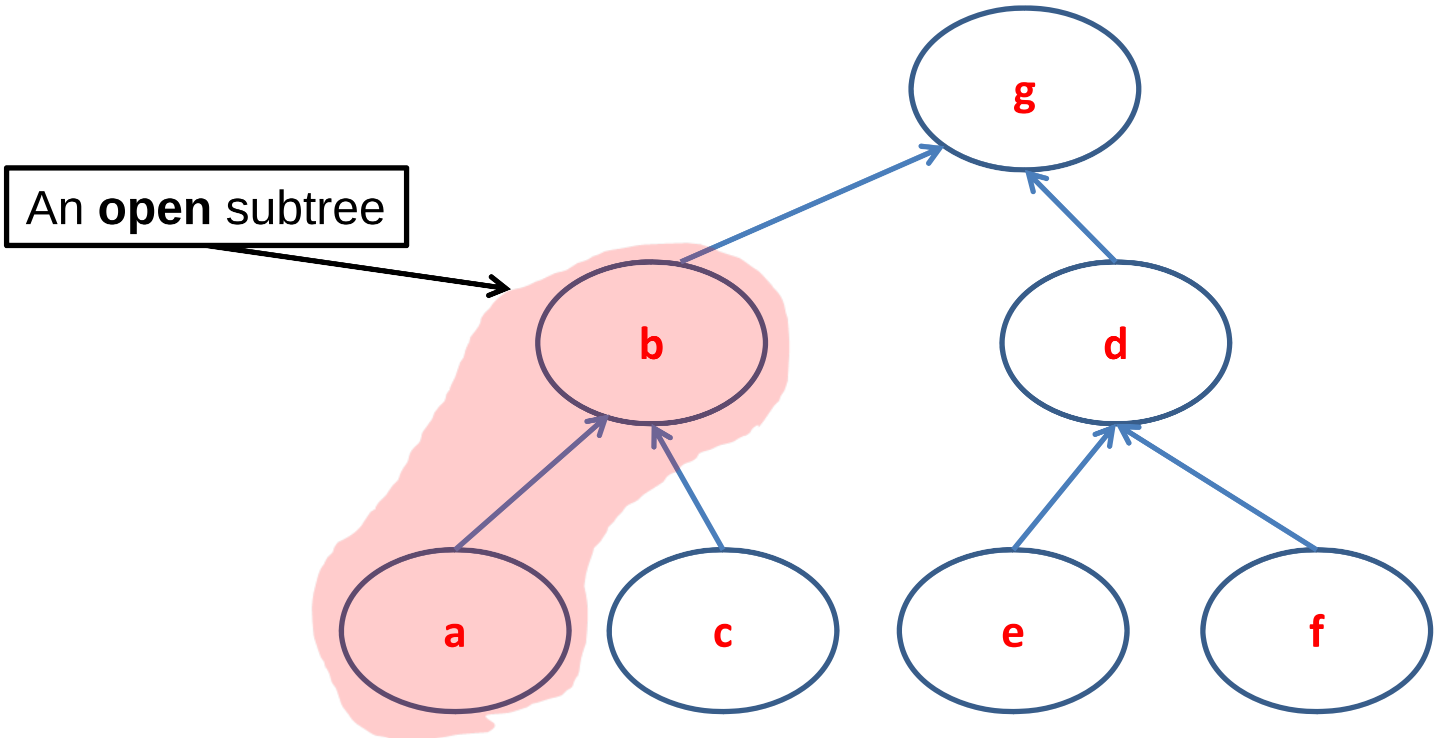
- Pros
  - intuitively model flat word movements
  - well-defined for phrase-based framework
- Cons
  - No linguistics structures
  - Need alignment matrix to determine movements

Completed/Open subtrees



All words under a node have been translated then we call a ***completed*** *subtree*

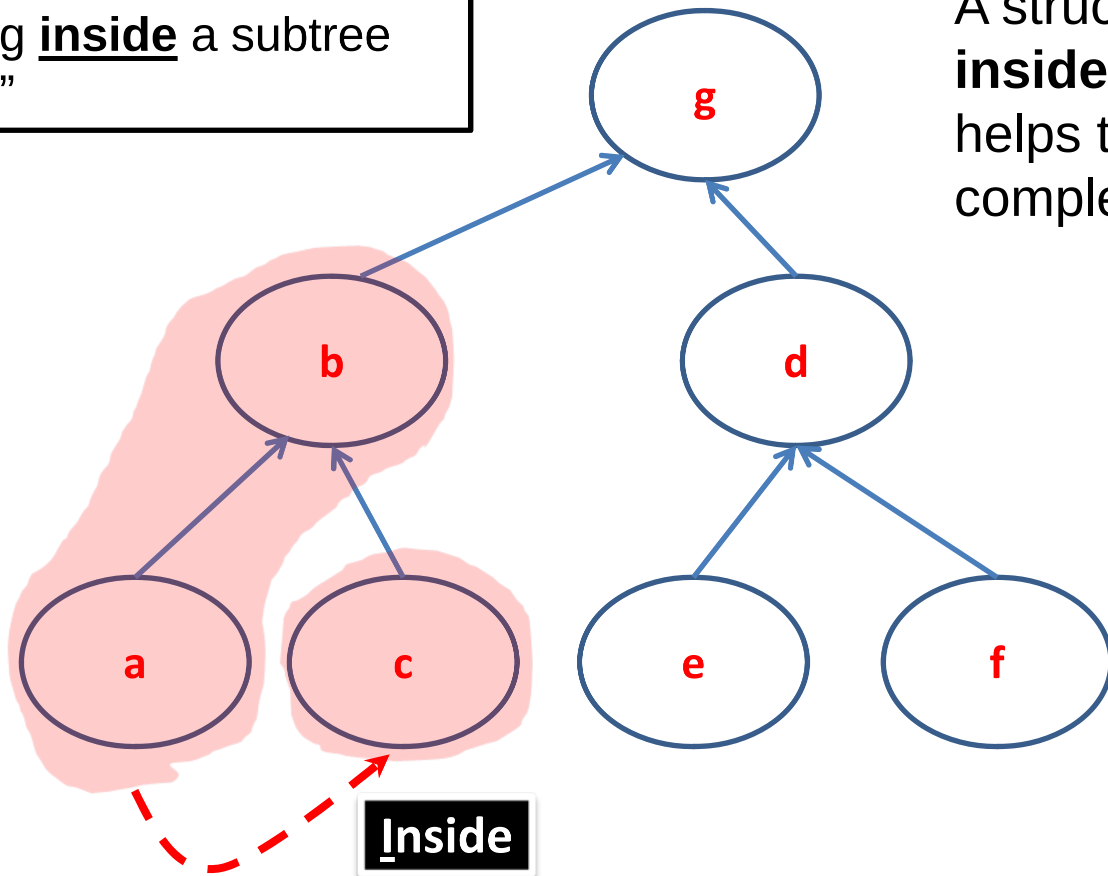
Completed/Open subtrees



A subtree that has begun translation but not yet complete, an ***open*** subtree

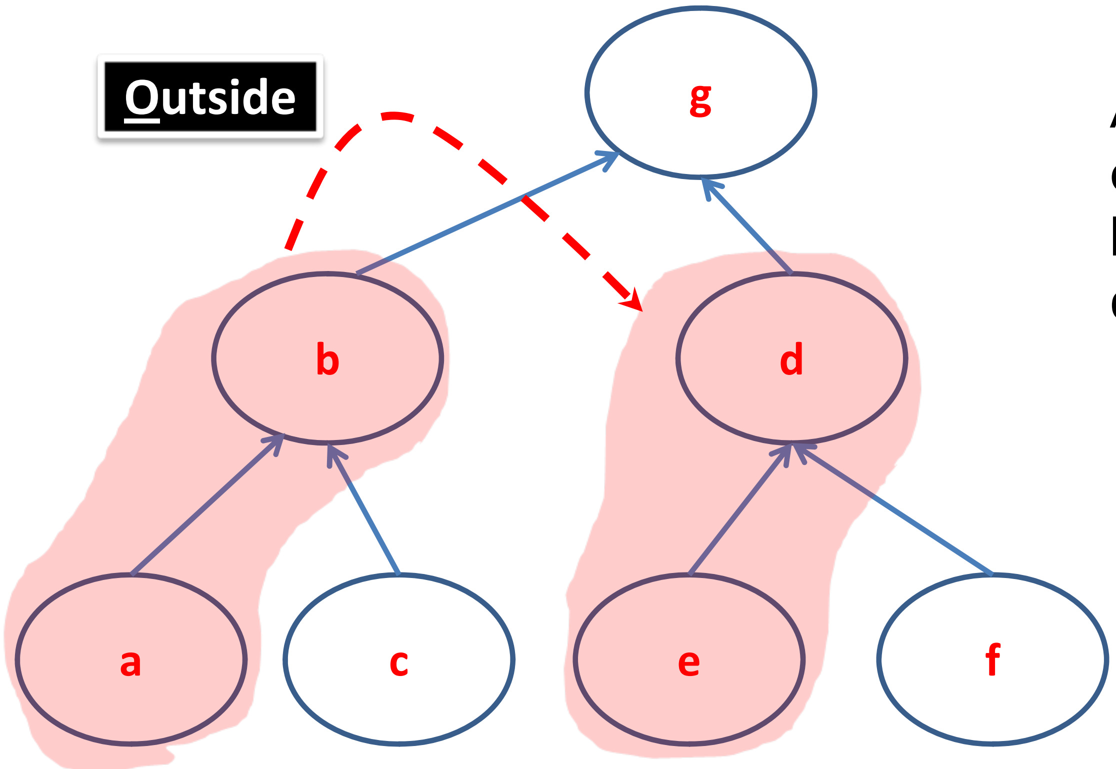
Inside/Outside subtree movements

“c” is moving **inside** a subtree rooted at “b”



A structure is moving **inside** a subtree if it helps the subtree to be completed or less open

Inside/Outside subtree movements



A structure is moving **outside** a subtree if it leaves the subtree to be open

"d e" is moving **<u>outside</u>** a subtree rooted at "b"

Source-side Dependency Tree (SDT) Reordering Models

$$p(D \mid e, f) = \prod_{i=1}^{n} p(d_i \mid \bar{e}_i, \bar{f}_{a_i}, a_i, s_{i-1}, s_i)$$

where

f is the input sentence;

$e = (\bar{e}_1, \ldots, \bar{e}_n)$ is the target language phrases;

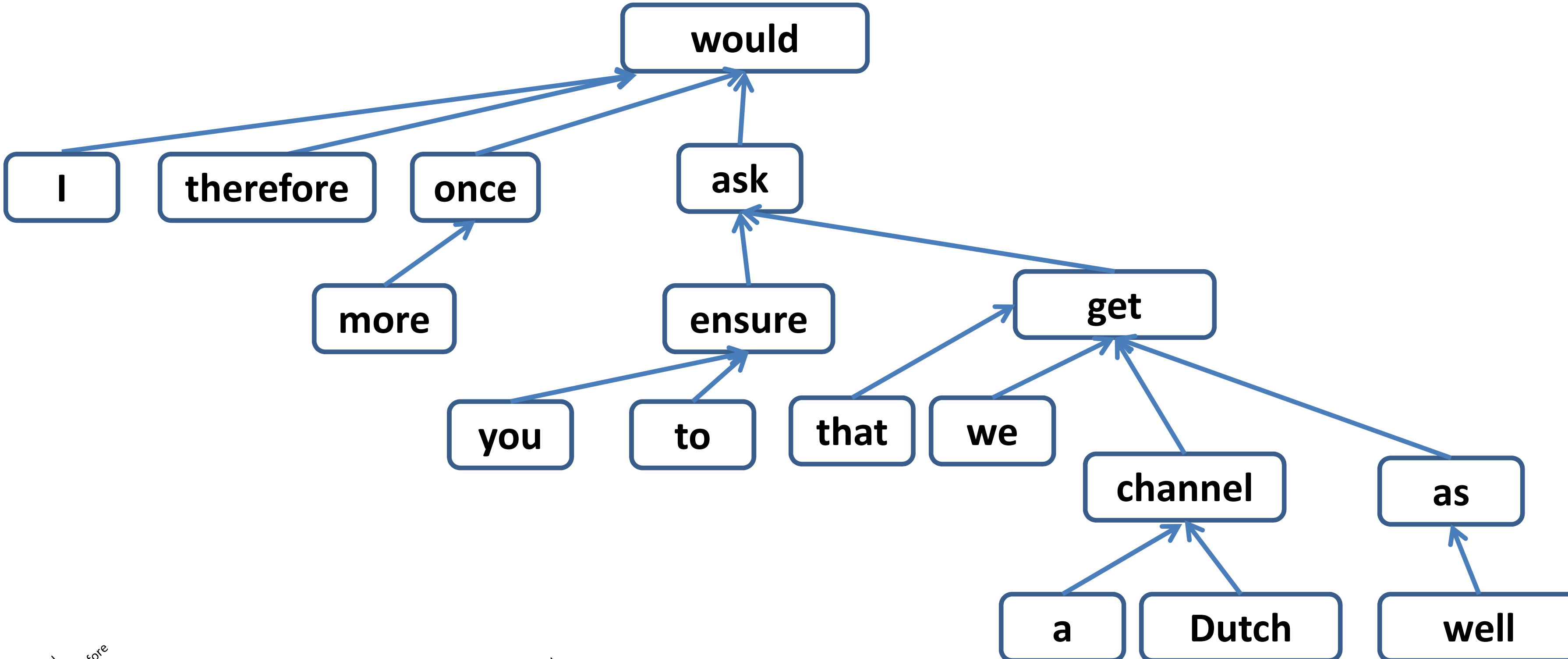
$a = (a_1, \ldots, a_n)$ is phrase alignments ;

$\bar{f}_{a_i}$ is a source phrase which has a translated phrase $\bar{e}_i$ defined by an alignment a_i;

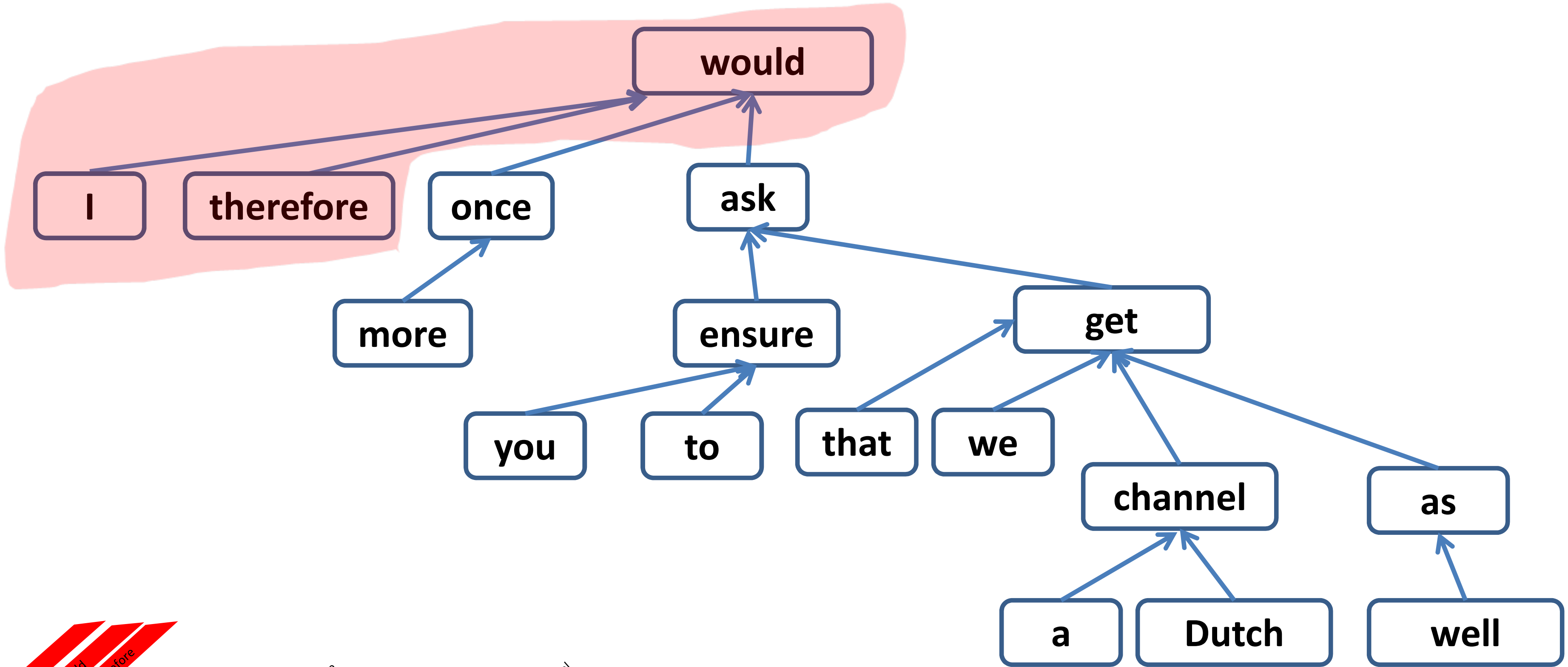
s_i and $s_{i\text{-}1}$ are dependency structures of source phrases $\bar{f}_{\mathrm{a_i}}$ and $\bar{f}_{\mathrm{a_{i\text{-}1}}}$;

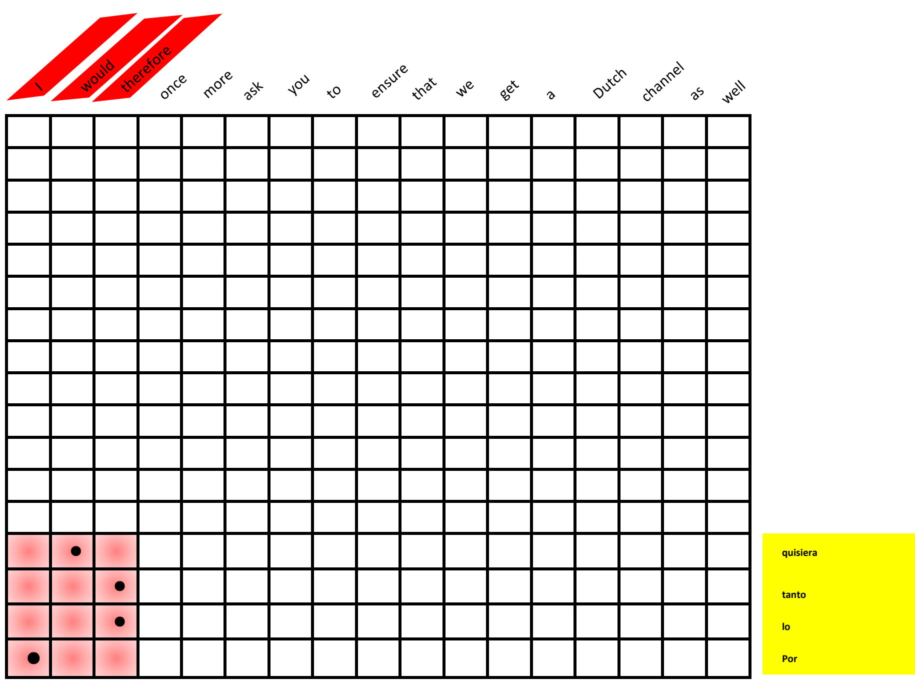
D represents the sequence of syntactic phrase movements over source dependency tree;

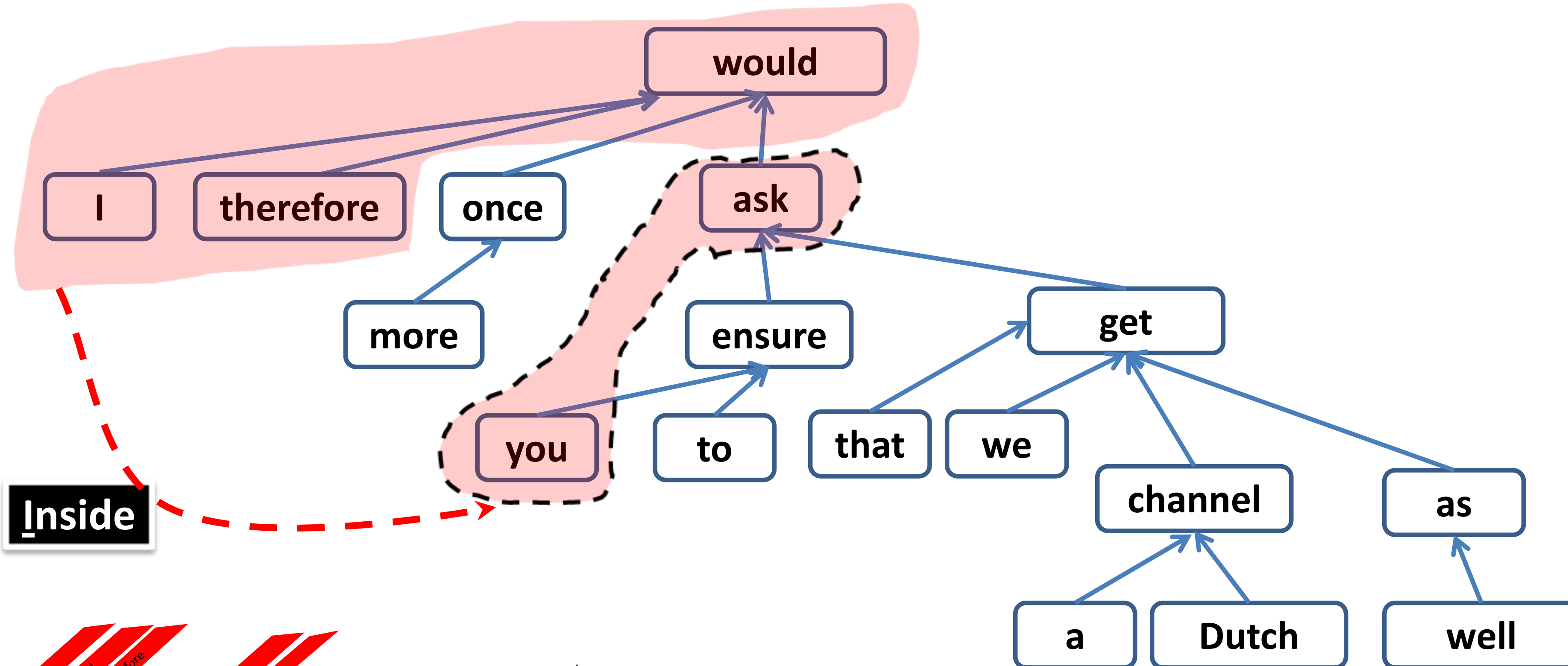
each $d_i = \{I, O\}$;



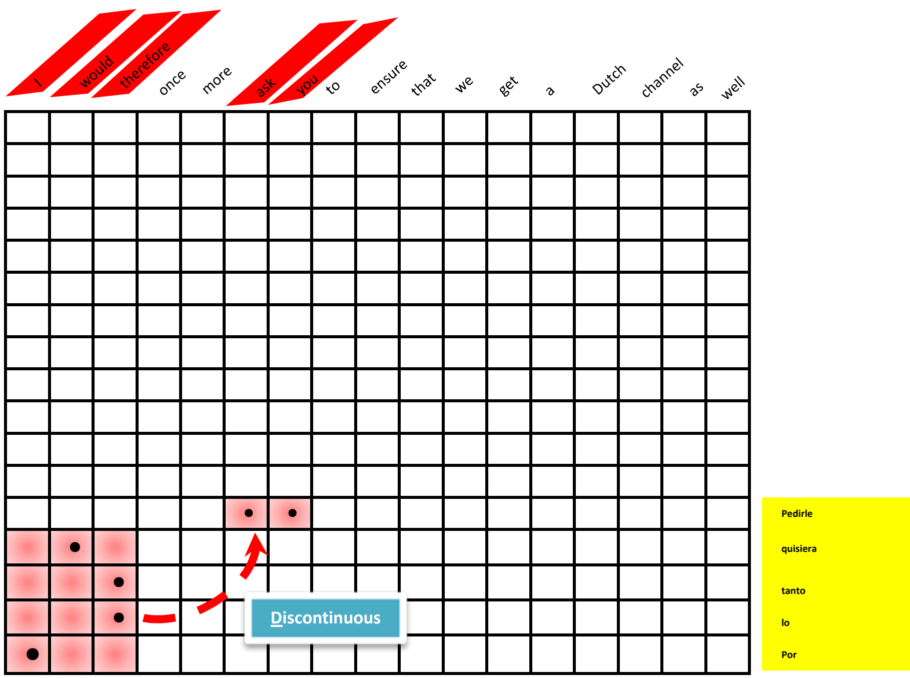
I	would	therefore	once	more	ask	you	to	ensure	that	we	get	a	Dutch	channel	as	well


would
I
therefore
once
ask
more
ensure
get
you
to
that
we
channel
as
a
Dutch
well

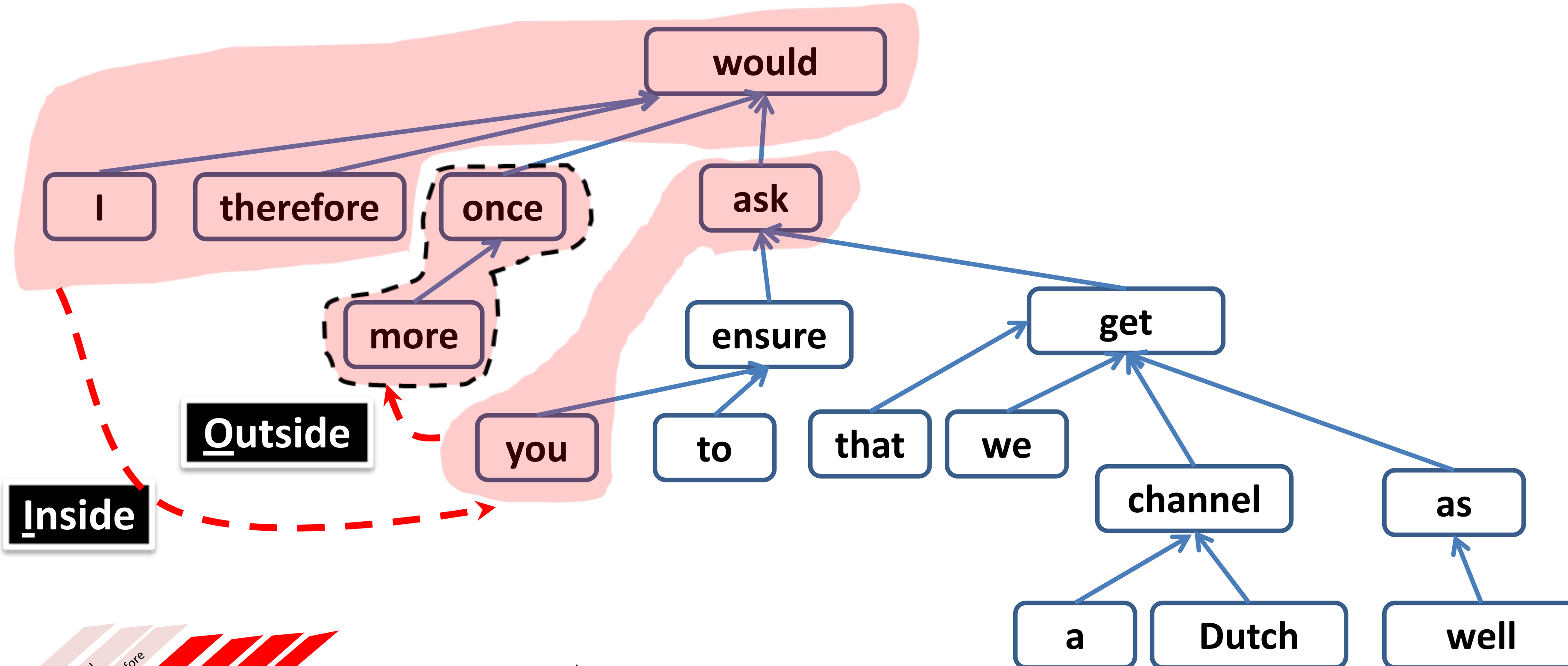

I
would
therefore
once
more
ask
you
to
ensure
that
we
get
a
Dutch
channel
as
well
quisiera
tanto
lo
Por

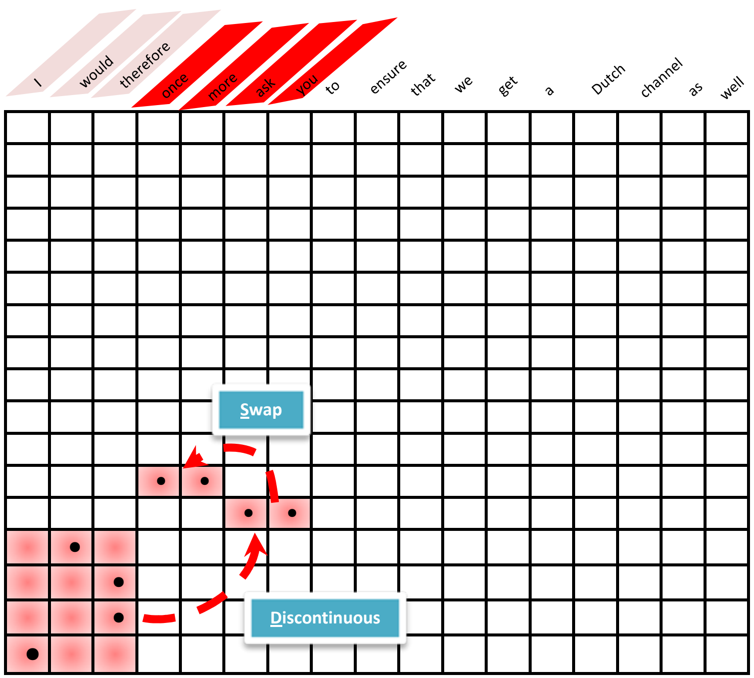


would
I
therefore
once
ask
more
ensure
get
you
to
that
we
channel
as
a
Dutch
well
Inside



I
would
therefore
once
more
ask
you
to
ensure
that
we
get
a
Dutch
channel
as
well
Discontinuous
Pedirle
quisiera
tanto
lo
Por





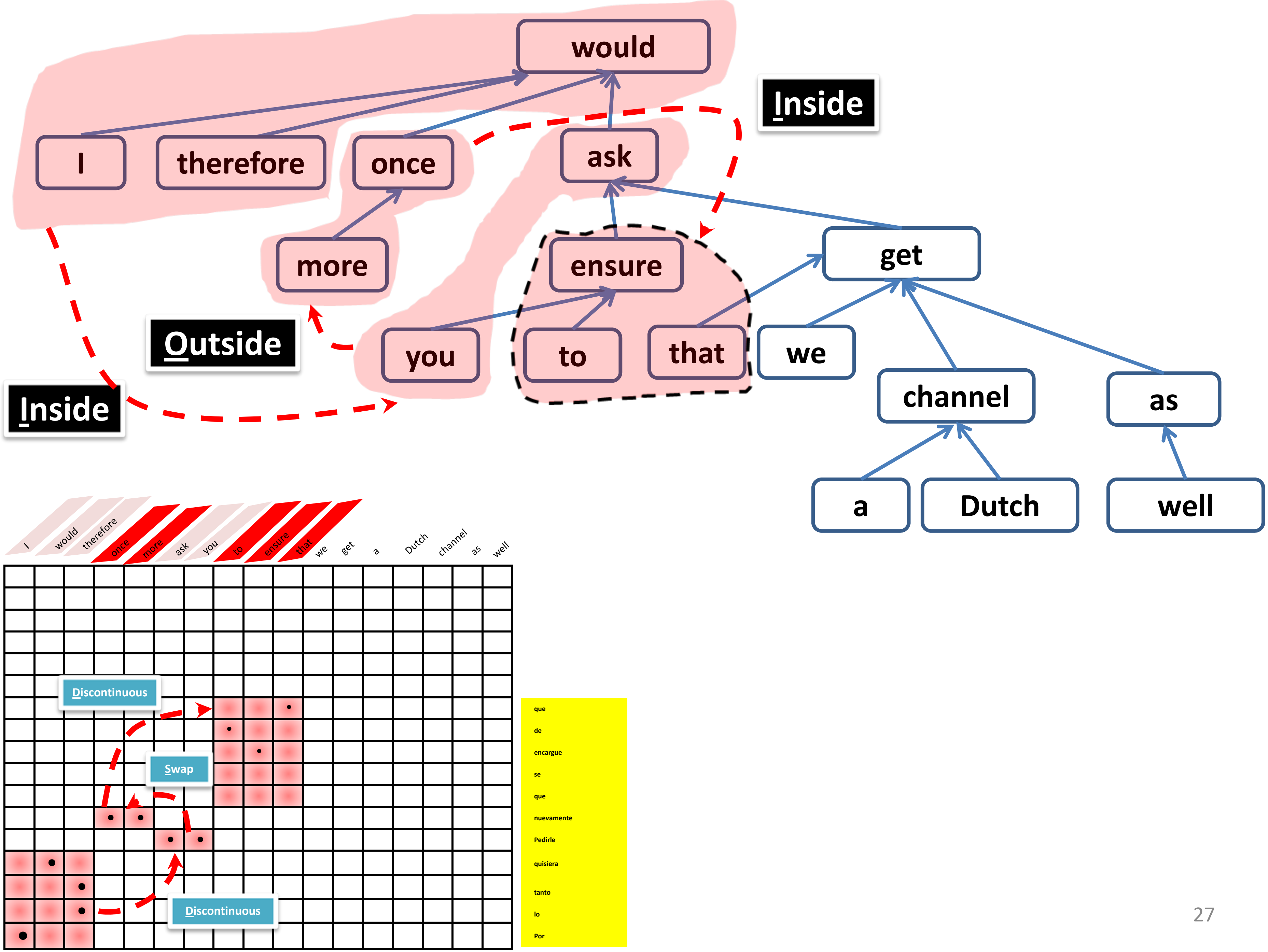
Pedirle

quisiera

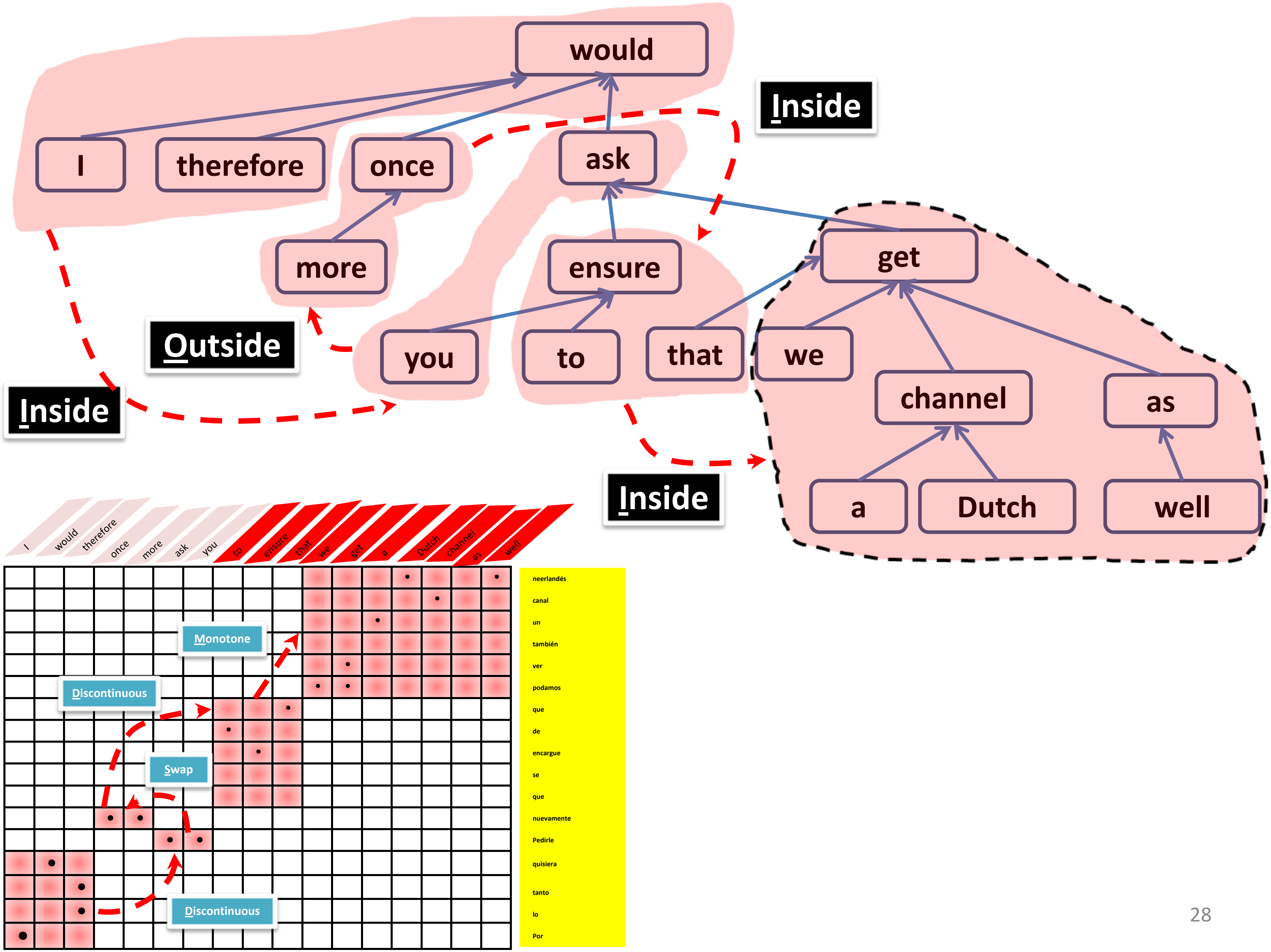
tanto

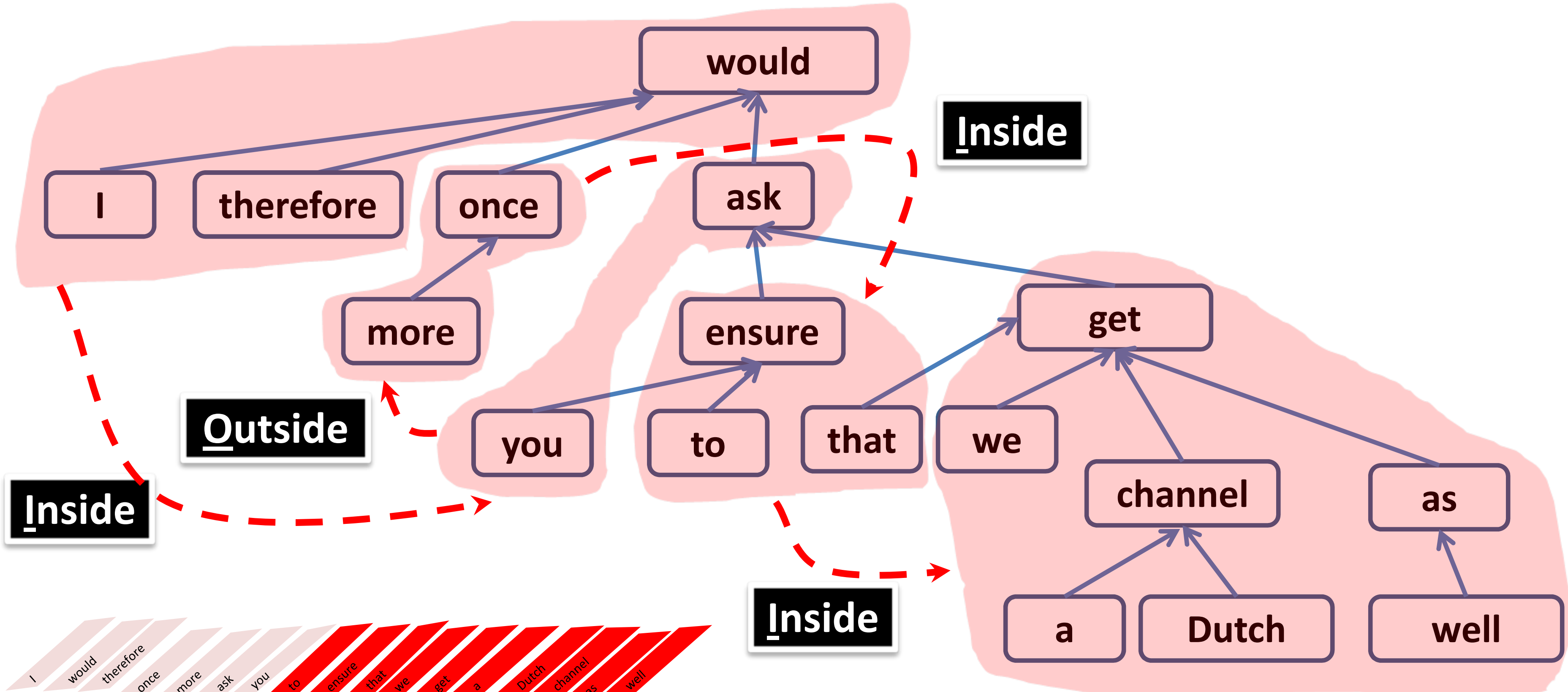
lo

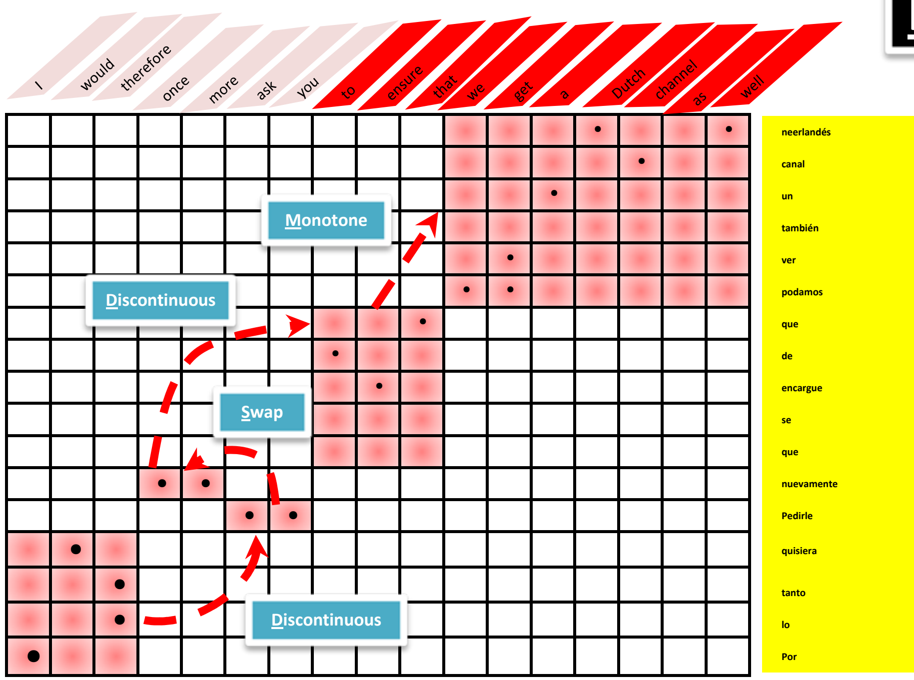
Por



would
Inside
I
therefore
once
ask
more
ensure
get
Outside
you
to
that
we
channel
as
Inside
a
Dutch
well
I
would
therefore
once
more
ask
you
to
ensure
that
we
get
a
Dutch
channel
as
well
Discontinuous
Swap
Discontinuous
que
de
encargue
se
que
nuevamente
Pedirle
quisiera
tanto
lo
Por


would
I
therefore
once
ask
more
ensure
you
to
that
get
we
channel
as
a
Dutch
well
Inside
Outside
Inside
Inside
Monotone
Discontinuous
Swap
Discontinuous
neerlandés
canal
un
también
ver
podamos
que
de
encargue
se
que
nuevamente
Pedirle
quisiera
tanto
lo
Por





Source-side Dependency Tree R.M.

Inside | Outside | Inside | Inside

Lexicalized R.M.

Discontinuous | Swap | Discontinuous | Monotone

Extended Source-side Dependency Tree (SDT) Reordering Models

$$p(D \mid e, f) = \prod_{i=1}^{n} p((o\_d)_i \mid \bar{e}_i, \bar{f}_{a_i}, a_i, a_{i-1}, s_{i-1}, s_i)$$

where

f is the input sentence;

$e = (\bar{e}_1, ..., \bar{e}_n)$ is the target language phrases;

$a = (a_1, ..., a_n)$ is phrase alignments ;

$\bar{f}_{a_i}$ is a source phrase which has a translate d phrase $\bar{e}_i$ defined by an alignment a_i;

s_i and s_{i-1} are dependency structures of source phrases $\bar{f}_{a_i}$ and $\bar{f}_{a_{i-1}}$;

D represents the sequence of syntactic phrase movements over source dependency tree;

each $(o\_d)_i = \{M\_I, S\_I, D\_I, M\_O, S\_O, D\_O\}$;

Extended Source-side Dependency Tree (SDT) Reordering Models

$p($

where

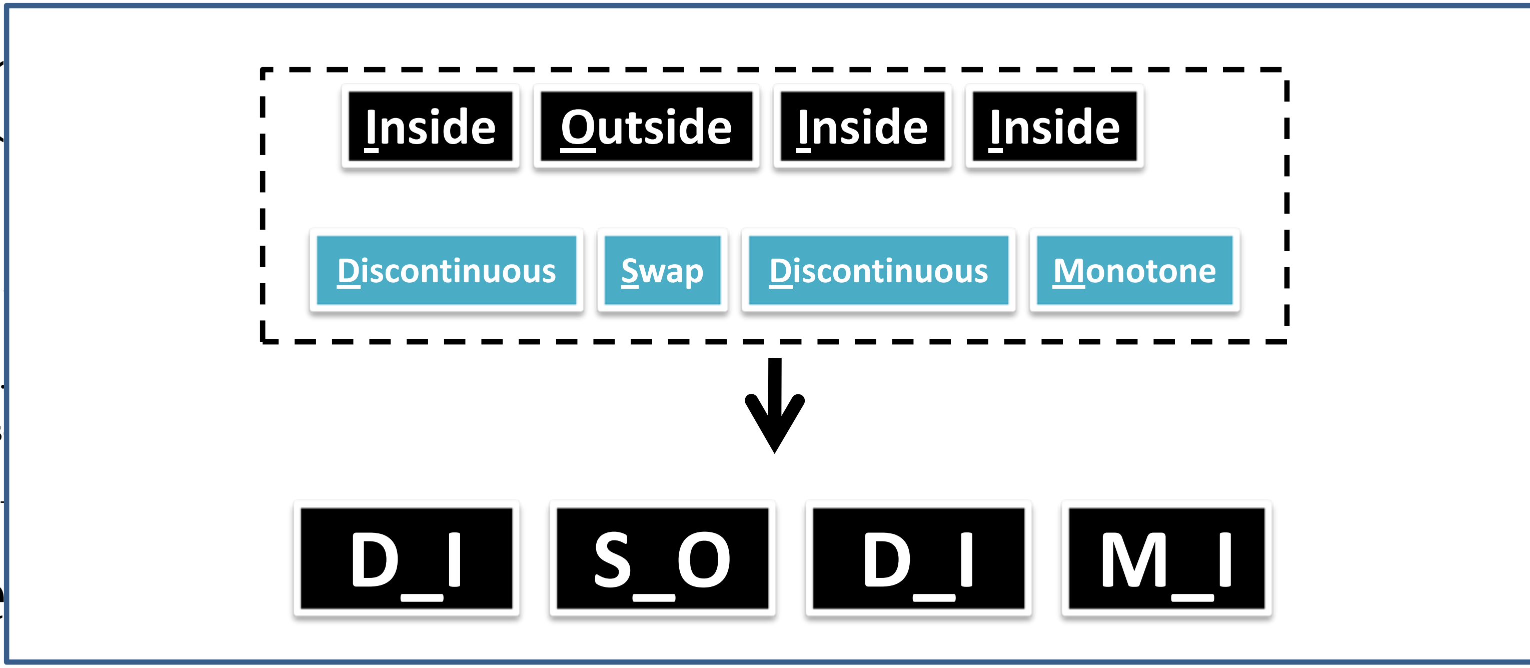
f is the

$e = (\bar{e}_1, \ldots$

$a = (a_1, \ldots$

$\bar{f}_{a_i}$ is a s

s_i and s_{i-}



D re ts

over source dependency tree;

each $(o\_d)_i = \{M\_I, S\_I, D\_I, M\_O, S\_O, D\_O\}$;

Training

- Obtain dependency parse of the source side
- Given a sentence pair and the source side dependency tree
  – Phrase extraction: also extract source dependency structures of phrase pairs
  – Identify Inside/Outside movement by using *Interruption Check Algorithms* (Bach et.al., 2009)

Training

DO: a **joint probability** of subtree movements and lexicalized orientations

$$p((o_j\_d_k) \mid \overline{e_i}, \bar{f}_{a_i}, o_j, d_k) = \frac{count(o_j\_d_k) + \gamma}{\sum_k \sum_j (count(o_j\_d_k) + \gamma)}$$

DOD: conditioned on subtree movements

$$p((o_j\_d_k) \mid \overline{e_i}, \bar{f}_{a_i}, d_k) = \frac{count(o_j\_d_k) + \gamma}{\sum_k (count(o_j\_d_k) + \gamma)}$$

DOO: conditioned on lexicalized orientations

$$p((o_j\_d_k) \mid \overline{e_i}, \bar{f}_{a_i}, o_j) = \frac{count(o_j\_d_k) + \gamma}{\sum_j (count(o_j\_d_k) + \gamma)}$$

Decoding

- Without cohesive constraints
  - Having no information about the source dependency tree information during the decoding time
  - Consider **both** subtree movements, and add them up to the translation model costs

- With cohesive constraints
  - The source dependency tree is available during the decoding time
  - Only consider **either** inside or outside movement, depending on the output of the interruption check algorithm

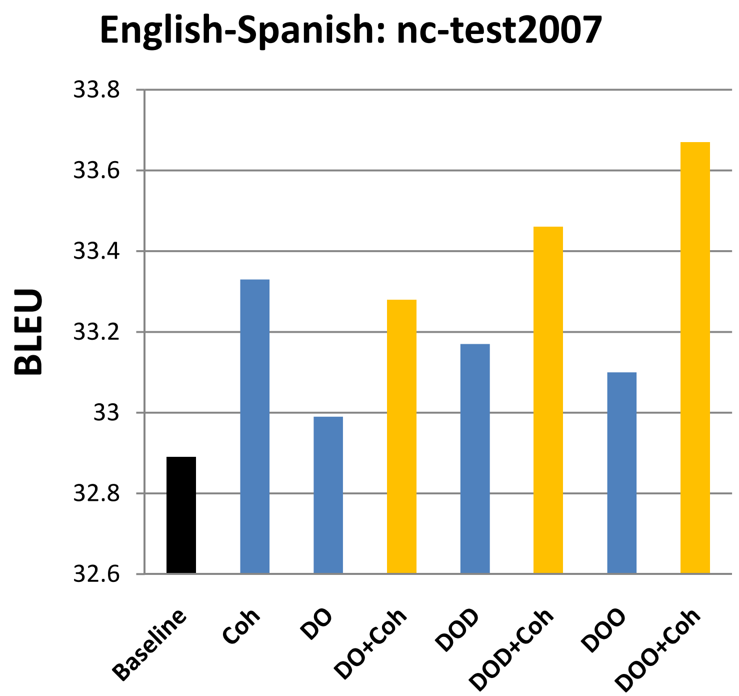
Outline

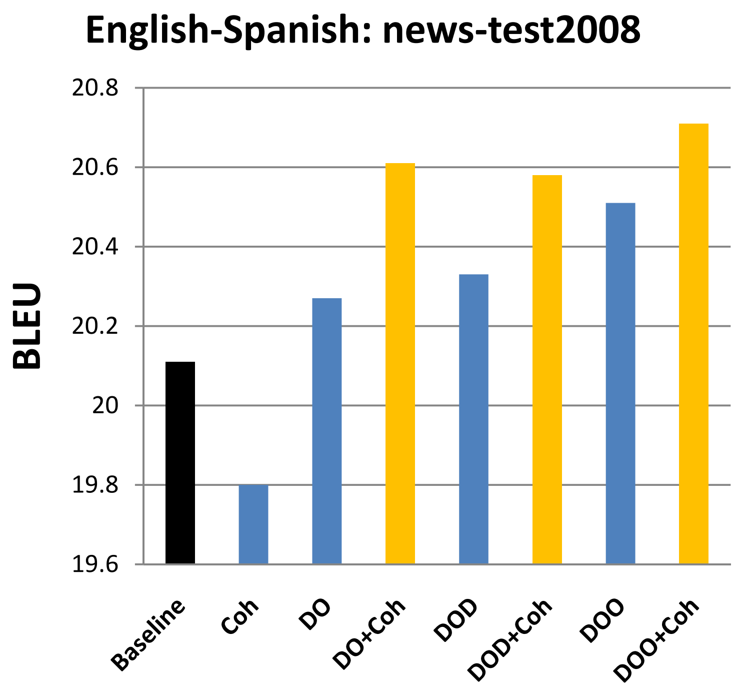
- Background of Reordering Models
- Source-side dependency tree reordering models
  - Modeling
  - Training
  - Decoding
- **Experiments & Analysis**
- Conclusion

Experiments setups

- **Baseline**: a phrase-based MT with lexicalized reordering model
- **Coh**: using cohesive constraints
- **DO** / **DOD** / **DOO**: using source-side dependency tree (SDT) reordering model with different parameter estimations
- **DO+Coh** / **DOD+Coh** / **DOO+Coh**: decoding with both SDT reordering model and cohesive constraints.

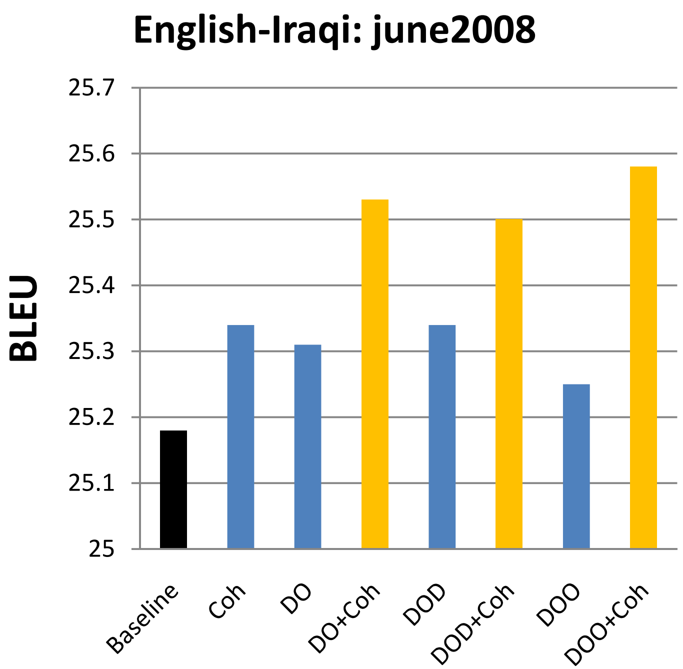
English-Spanish (Europarl)

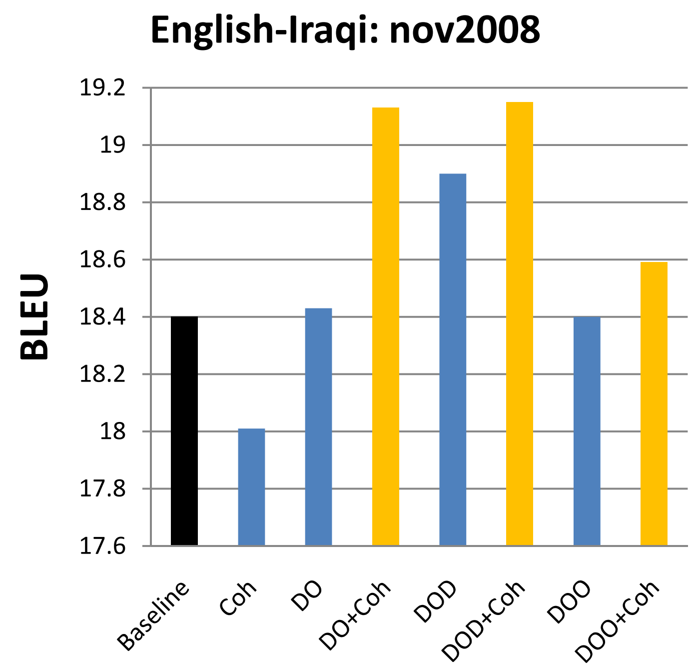




- Source-side dependency tree reordering models and cohesive constraints obtained **improvements** over the lexicalized reordering models.

English-Iraqi (TransTac)



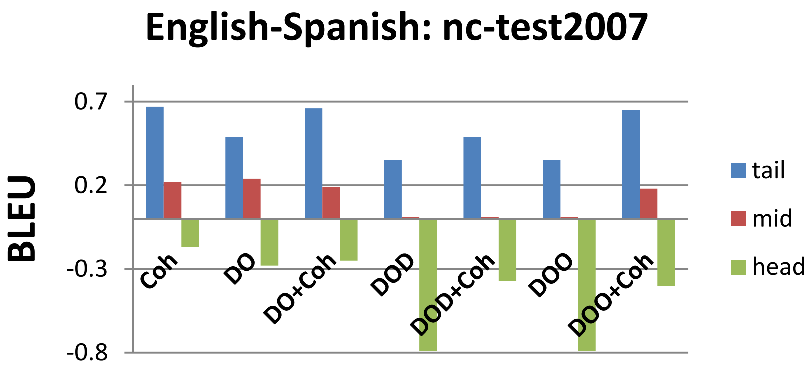


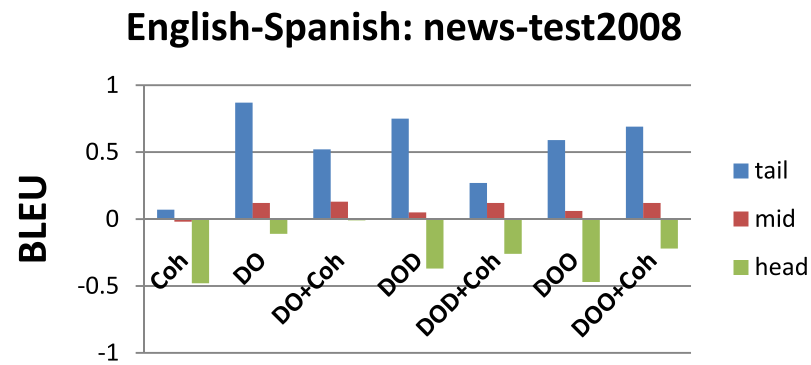
- Decoding with both source-side dependency tree reordering models and cohesive constraints often obtain the best performance.

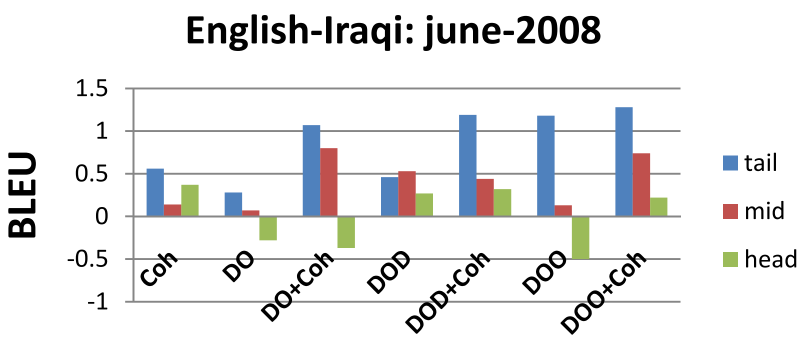
Where are improvements coming from?

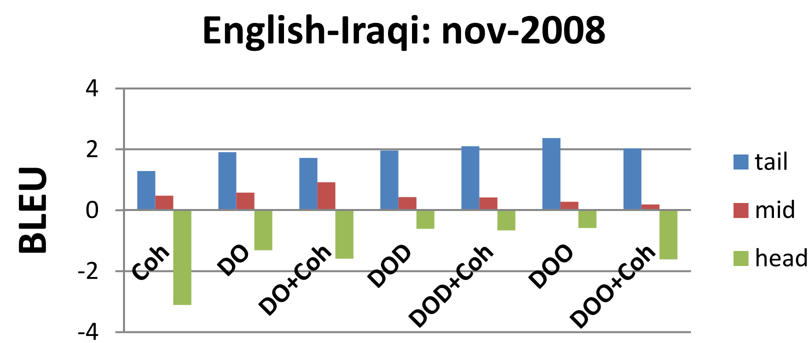
Test set breakdown

- Divide the test sets into three portions based on sentence-level TER of the baseline system
- μ and σ are mean and standard deviation of the whole test set
- **Head**, **Tail** and **Mid** as the sentence whose score is lower than μ-1/2 σ, higher than μ+1/2 σ and the rest









	june-08	nov-08	nc-test2007	news-test2008
Head	7.92	6.27	20.39	13.07
Mid	12.31	11.09	28.07	22.78
Tail	13.91	14.08	35.29	25.33

What is the most significant effect the source-tree reordering models contribute?

Numbers of Reorderings

	nc-test2007	news-test2008	june-2008	nov-2008
Baseline	1507	1684	39	24
Coh	2045	**2903**	46	21
DO	**2189**	2113	97	58
DO+Coh	1929	1900	155	88
DOD	1735	2592	123	60
DOD+Coh	2070	2021	148	**90**
DOO	1735	1785	164	49
DOO+Coh	1818	1959	**247**	66

- More reorderings can be generated without losing performance.
- The source-tree reordering models provide a more discriminative mechanism to estimate reordering events.
- Reordering is more language-specific than general translation models, and the conditions for a reordering event to happen vary among languages.

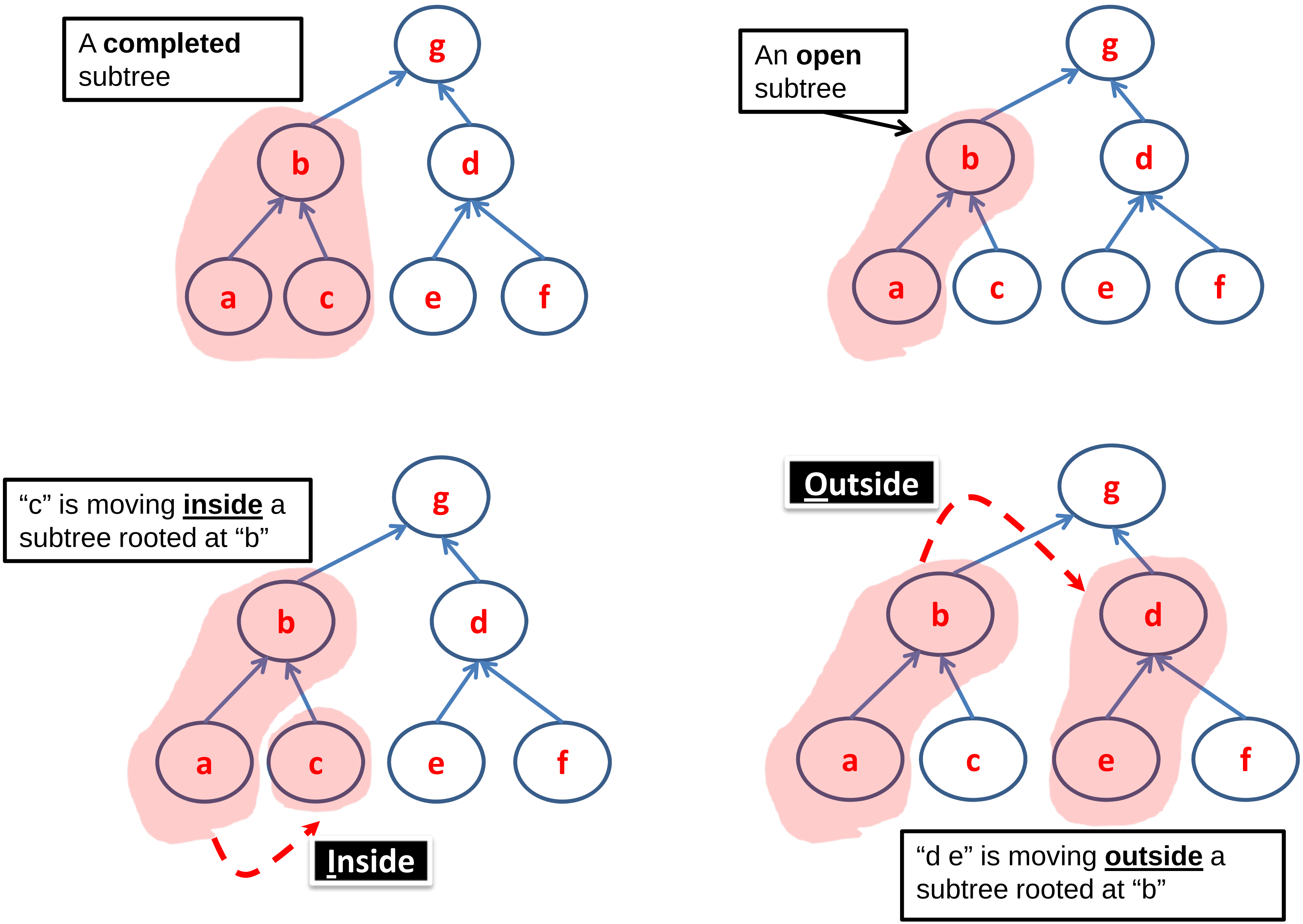
Outline

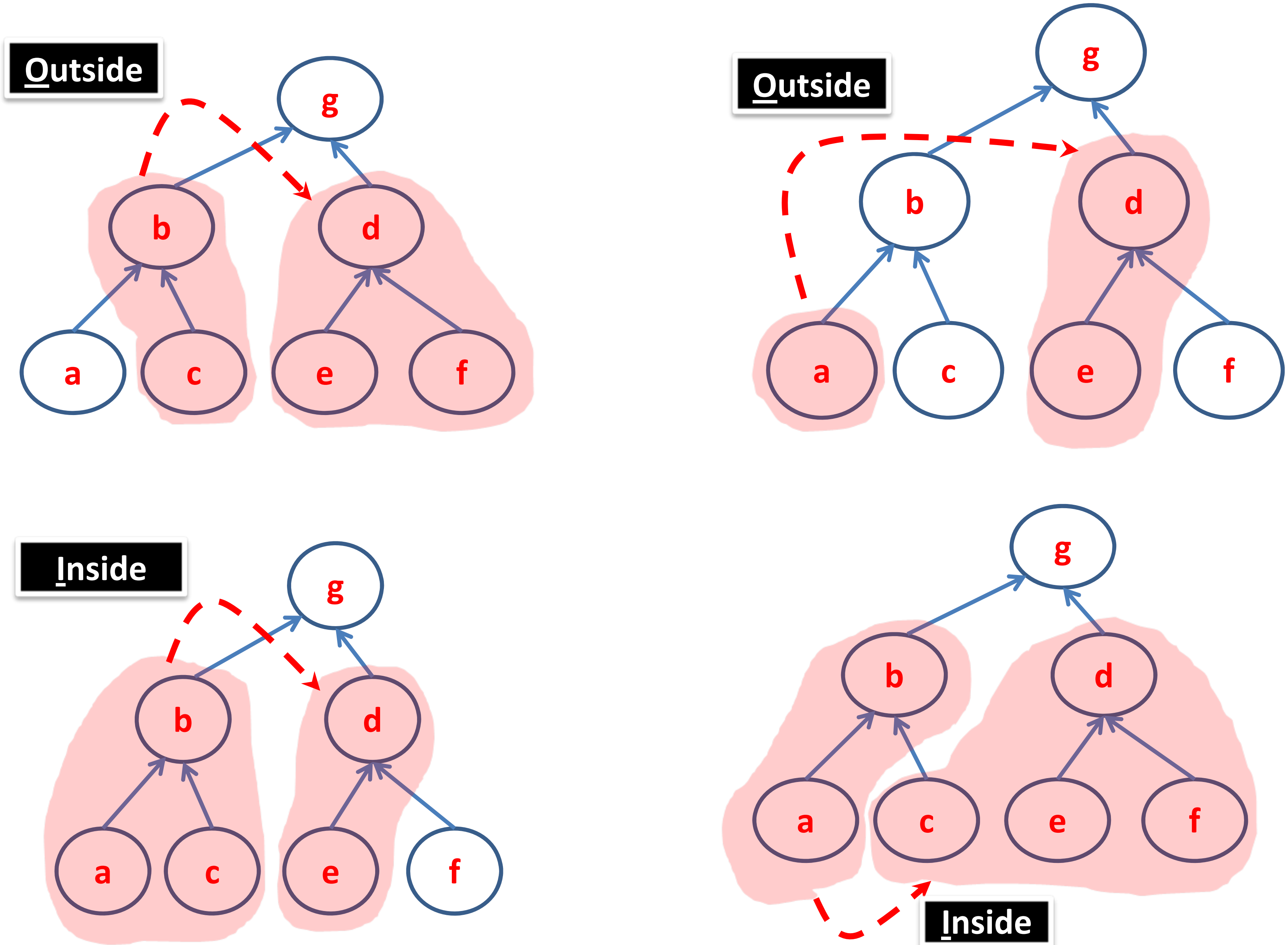
- Background & Motivations
- Source-side dependency tree reordering models
  - Modeling
  - Training
  - Decoding
- Experiments & Analysis
- **Conclusions**

Conclusions & Future Work

- Conclusions
  - Source-side dependency tree reordering models are helpful
    - Model reordering event with **Inside/Outside** subtree movements
  - The effectiveness was shown when comparing with a strong reordering model
  - Obtained improvements with 2 language pairs and also covered a training corpus sizes, ranging from 500K up to 1.3M sentence pairs
- Future work
  - A hierarchical source side dependency reordering model: extend Galley&Manning (2008).
  - Packed-forest dependency tree reordering models

Back up


A **completed** subtree
g
b
d
a
c
e
f
An **open** subtree
g
b
d
a
c
e
f
“c” is moving **inside** a subtree rooted at “b”
g
b
d
a
c
e
f
Inside
Outside
g
b
d
a
c
e
f
“d e” is moving **outside** a subtree rooted at “b”


Outside
g
b
d
a
c
e
f
Outside
g
b
d
a
c
e
f
Inside
g
b
d
a
c
e
f
g
b
d
a
c
e
f
Inside

What do you mean by introducing Inside/Outside notions?

- The movement of the **subtree inside** or **outside** a source subtree can be viewed as the decoder is leaving from the **previous** source state to the **current** source state.

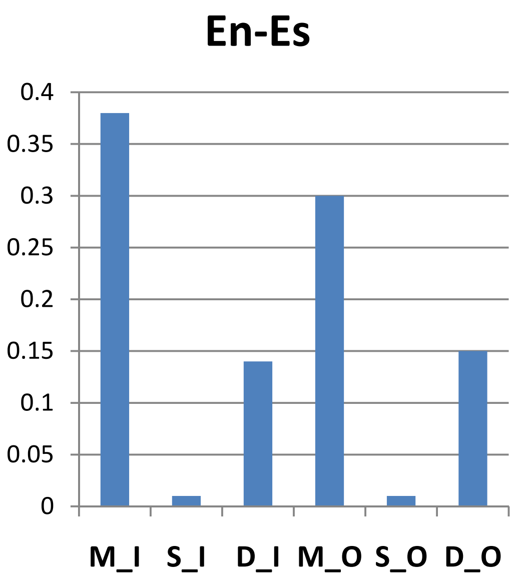
- Tracking facts about the **subtree-to-subtree transitions** observed in the source side of word-aligned training data.

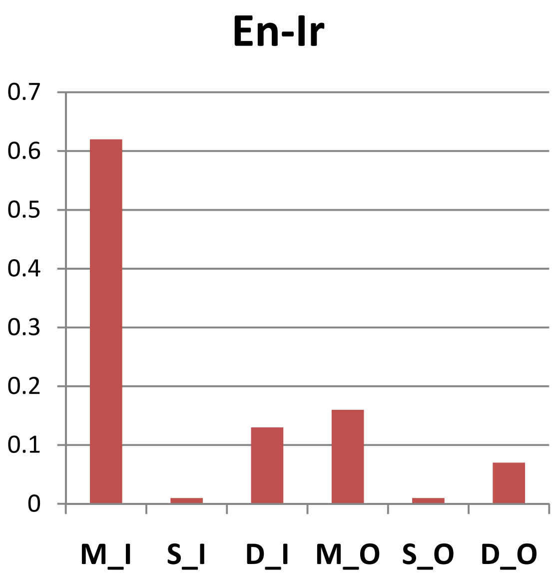
	Lexicalized	Source-tree
ask you # pedirle	dis swap	D_I *
ask you # pedirle	mono mono	M_I
ask you # pedirle	mono mono	M_O
once more # nuevamente	swap dis	S_O *
once more # nuevamente	dis swap	D_O
once more # nuevamente que	swap dis	S_O

inside and outside probabilities for phrase "***ask you***"- "***pedirle***" according to three parameter estimation methods

	M_I	S_I	D_I	M_O	S_O	D_O
DO	0.691	0.003	0.142	0.119	0.009	0.038
DOD	0.827	0.003	0.17	0.719	0.053	0.228
DOO	0.854	0.25	0.79	0.146	0.75	0.21

Distributions of Reordering Events





Observed **monotone** & **inside** (**M_I**) movements more often than other categories

Reordering Models

- Explicitly model phrase reordering distances
  - Distance-based (Och, 2002; Koehn et.al., 2003)
  - Lexicalized phrase (Tillmann, 2004; Koehn, et.al., 2005; Al-Onaizan and Papineni, 2006)
  - Hierarchical phrase (Galley and Manning, 2008)
  - MaxEnt classifier (Zens and Ney, 2006; Xiong, et.al., 2006; Chang, et. al., 2009)
- Put syntactic analysis of the target language into both modeling and decoding
- Use source language syntax

Reordering Models

- Explicitly model phrase reordering distances
  - Distance-based (Och, 2002; Koehn et.al., 2003)
  - Lexicalized phrase (Tillmann, 2004; Koehn, et.al., 2005; Al-Onaizan and Papineni, 2006)
  - Hierarchical phrase (Galley and Manning, 2008)
  - MaxEnt classifier (Zens and Ney, 2006; Xiong, et.al., 2006; Chang, et. al., 2009)
- Put syntactic analysis of the target language into both modeling and decoding
  - Direct model target language constituents movement in either constituency trees (Yamada and Knight, 2001; Galley et.al., 2006; Zollmann et.al., 2008) or dependency trees (Quirk, et.al., 2005)
  - Hierarchical phrase-based (Chiang, 2005; Shen et. al., 2008)
- Use source language syntax