Disambiguation Strategies for Data-Oriented Translation

Mary Hearne and Andy Way

National Centre for Language Technology
Dublin City University

June 14, 2006

Outline

Outline

Data-Oriented Translation (DOT)

Data-Oriented Translation is:

a hybrid model of translation which combines **examples**, **linguistic information** and a **statistical translation model**.

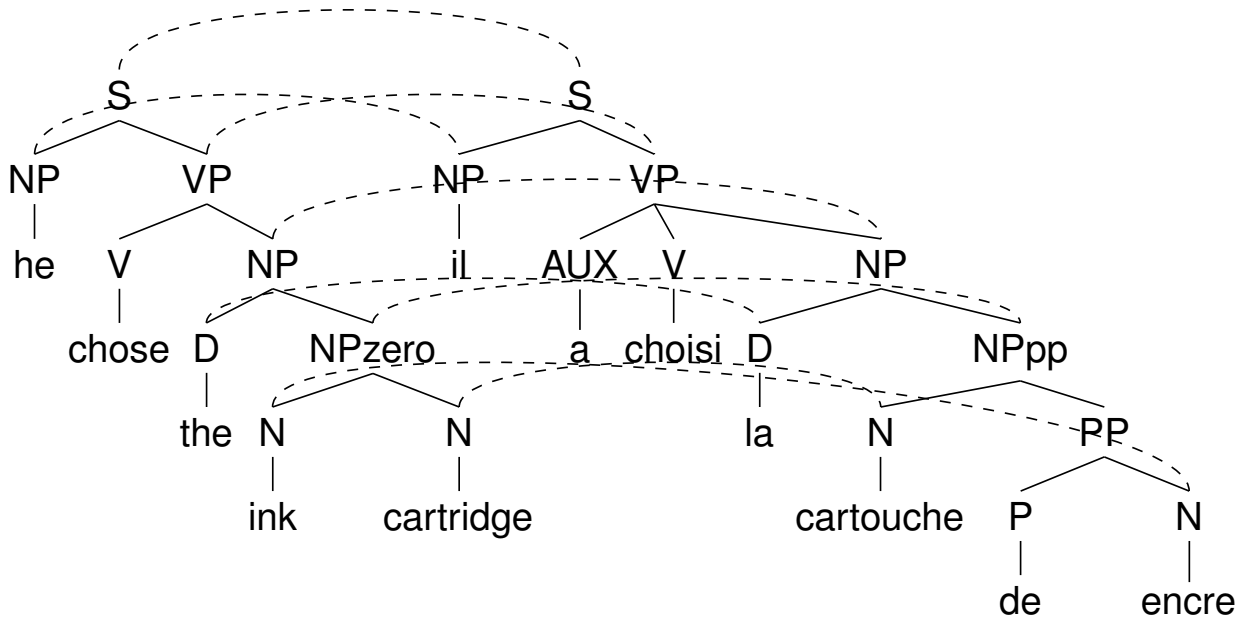
The DOT model is specified in terms of:

- the type of representation expected in the example base,
- how fragments are to be extracted from these representations,
- how extracted fragments are to be recombined when analysing and translating input sentences, and
- how the resulting translations are to be ranked.

DOT: representations

Tree-DOT representations comprise:

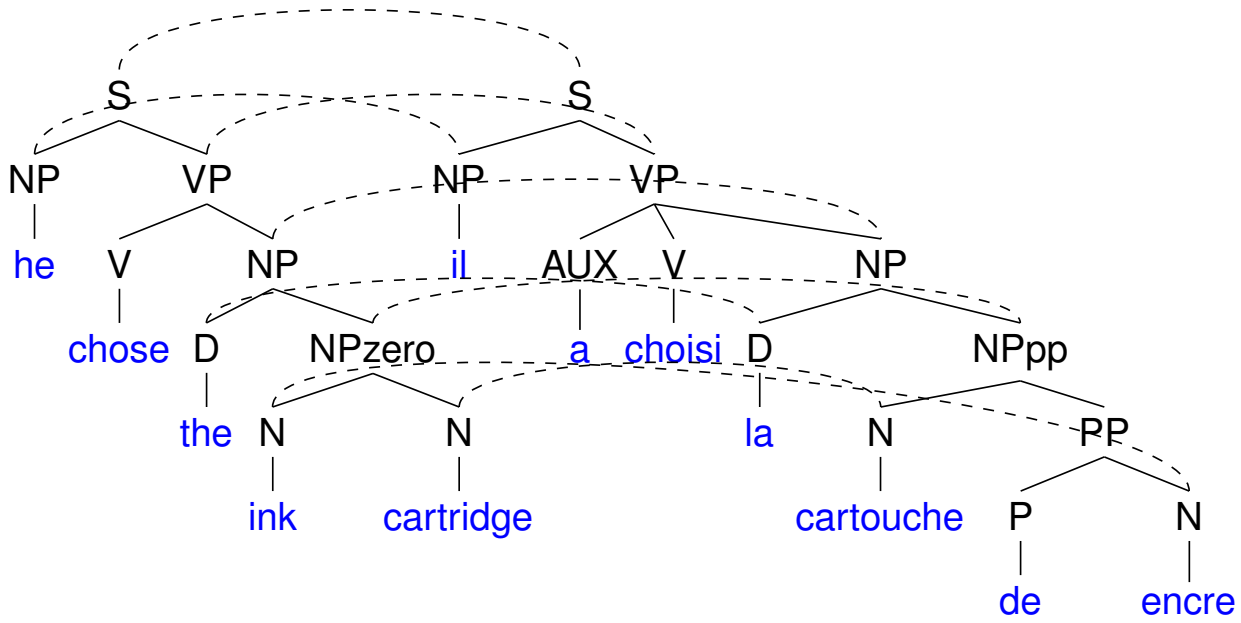
- aligned sentence pairs
- annotated with context-free phrase-structure trees
- sub-structural links denoting translational equivalence



DOT: representations

Tree-DOT representations comprise:

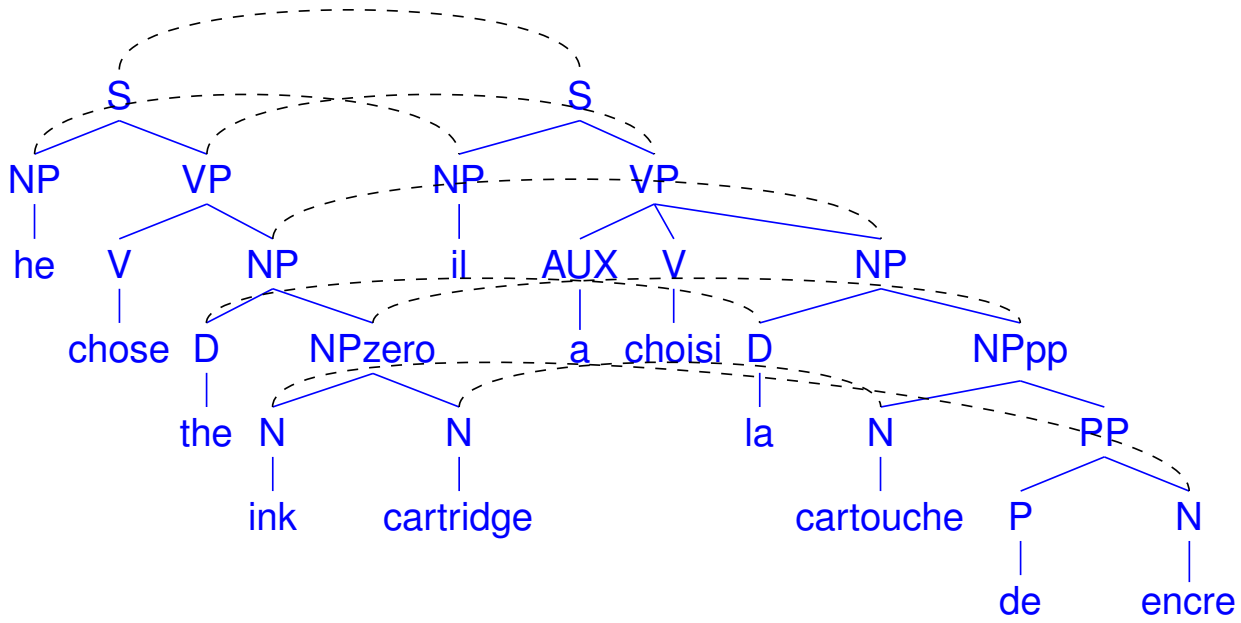
- aligned sentence pairs
- annotated with context-free phrase-structure trees
- sub-structural links denoting translational equivalence



DOT: representations

Tree-DOT representations comprise:

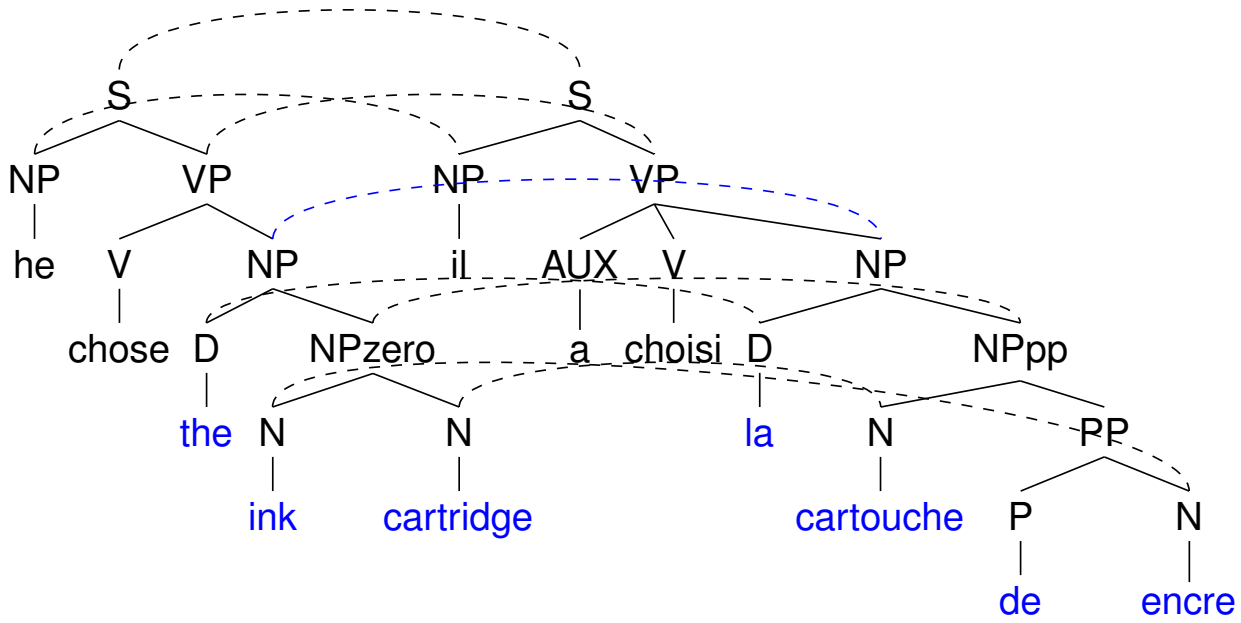
- aligned sentence pairs
- annotated with context-free phrase-structure trees
- sub-structural links denoting translational equivalence



DOT: representations

Tree-DOT representations comprise:

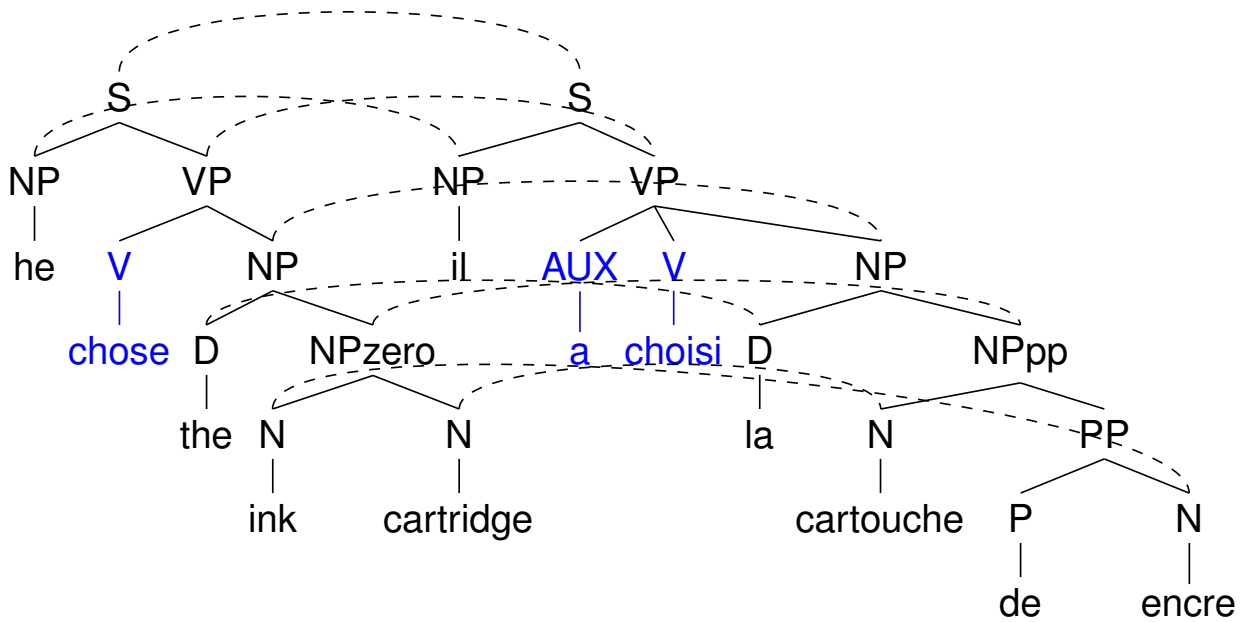
- aligned sentence pairs
- annotated with context-free phrase-structure trees
- sub-structural links denoting translational equivalence



DOT: representations

Tree-DOT representations comprise:

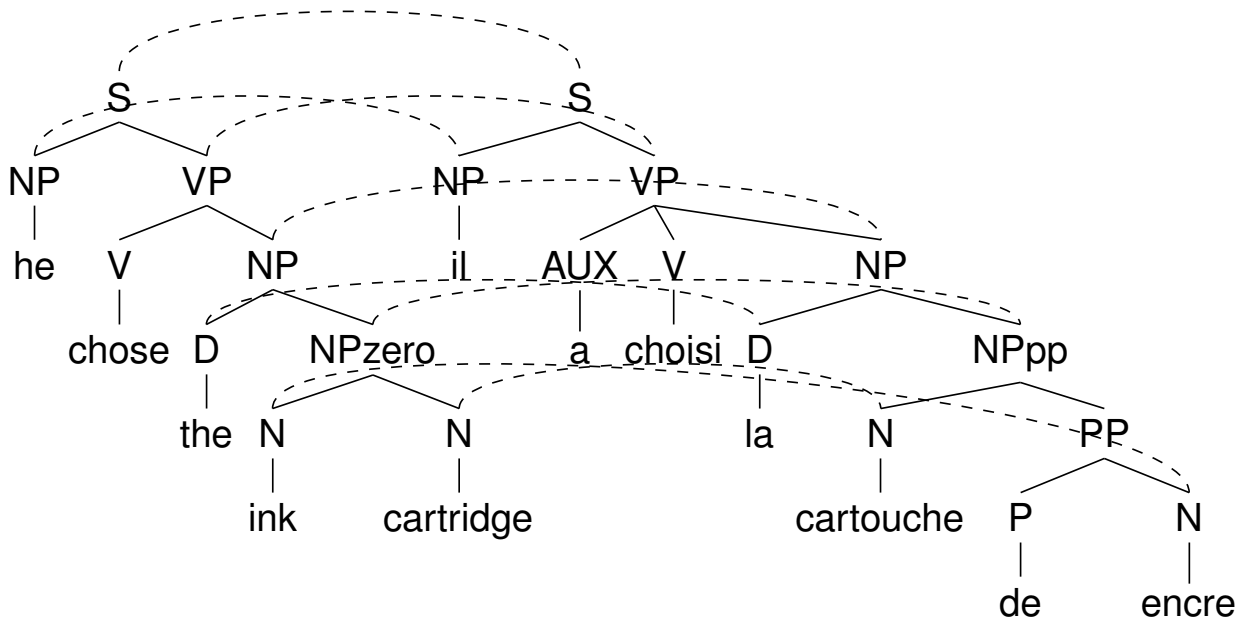
- aligned sentence pairs
- annotated with context-free phrase-structure trees
- sub-structural links denoting translational equivalence



DOT: fragmentation

The *root* and *frontier* operations:

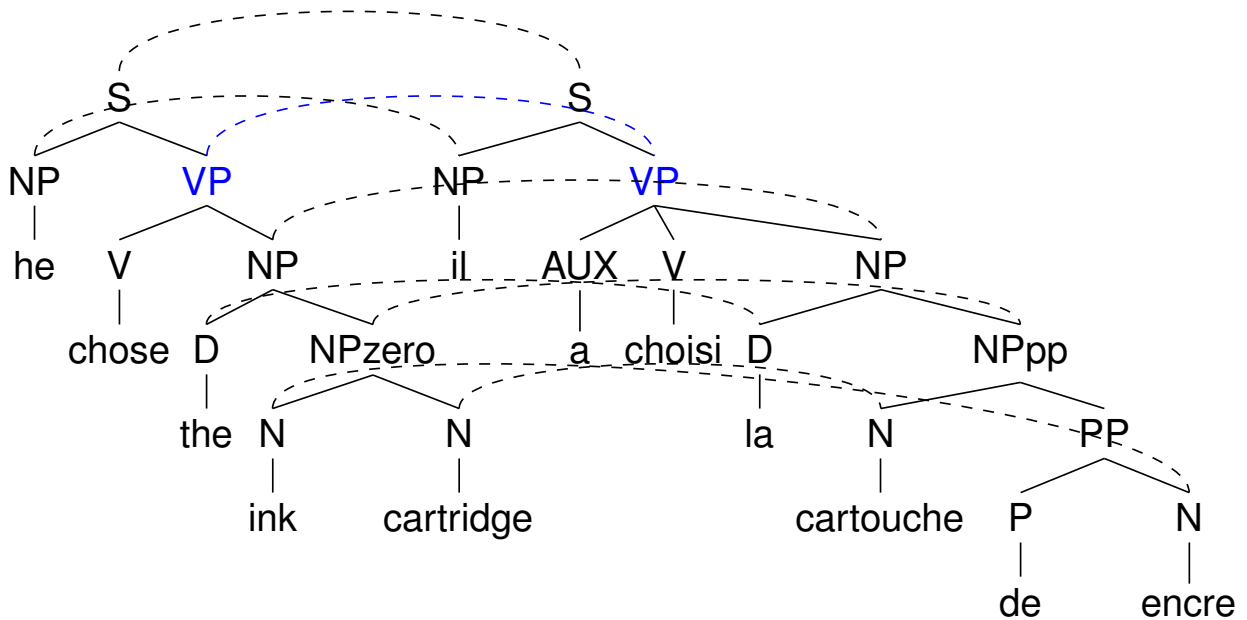
- select a **linked** node pair to be *root* nodes and delete all except these nodes, the subtrees they dominate and the links between them, and
- select a set of **linked** node pairs to be *frontier* nodes and delete the subtrees they dominate.



DOT: fragmentation

The *root* and *frontier* operations:

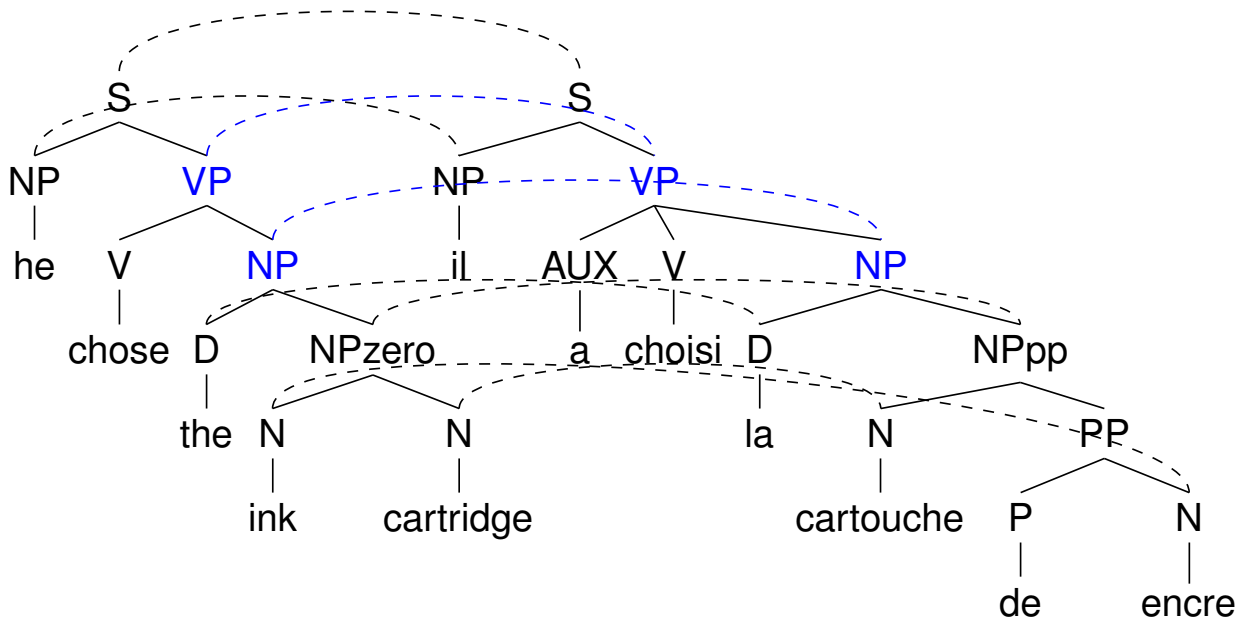
- select a **linked** node pair to be *root* nodes and delete all except these nodes, the subtrees they dominate and the links between them, and
- select a set of **linked** node pairs to be *frontier* nodes and delete the subtrees they dominate.



DOT: fragmentation

The *root* and *frontier* operations:

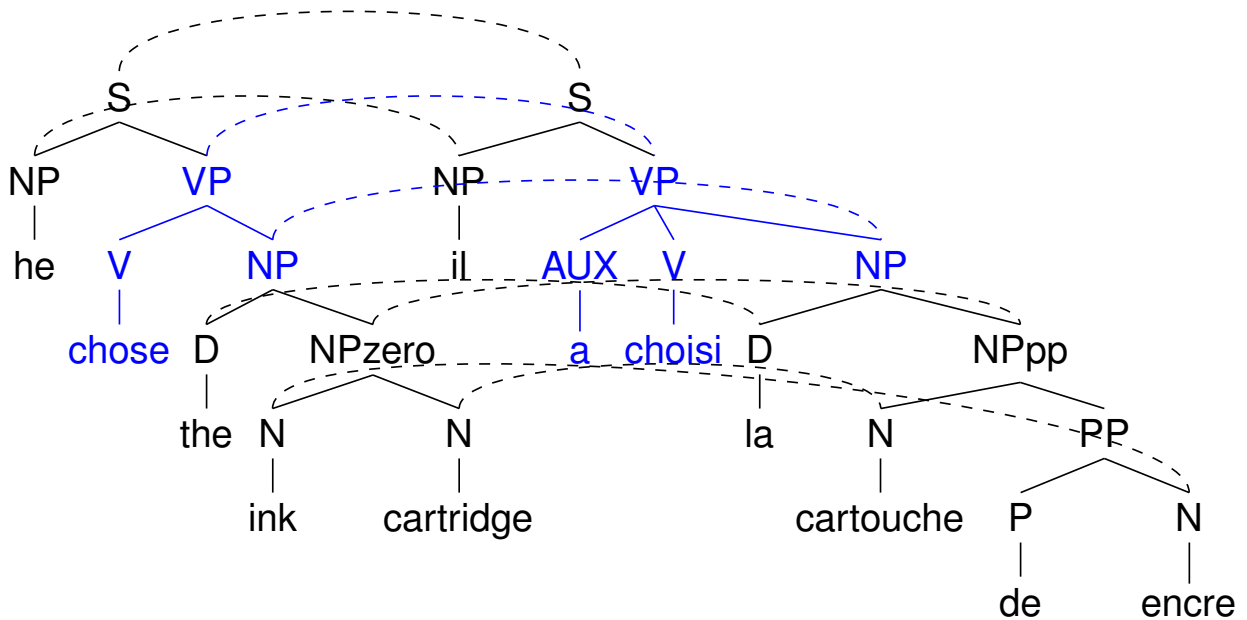
- select a **linked** node pair to be *root* nodes and delete all except these nodes, the subtrees they dominate and the links between them, and
- select a set of **linked** node pairs to be *frontier* nodes and delete the subtrees they dominate.



DOT: fragmentation

The *root* and *frontier* operations:

- select a **linked** node pair to be *root* nodes and delete all except these nodes, the subtrees they dominate and the links between them, and
- select a set of **linked** node pairs to be *frontier* nodes and delete the subtrees they dominate.



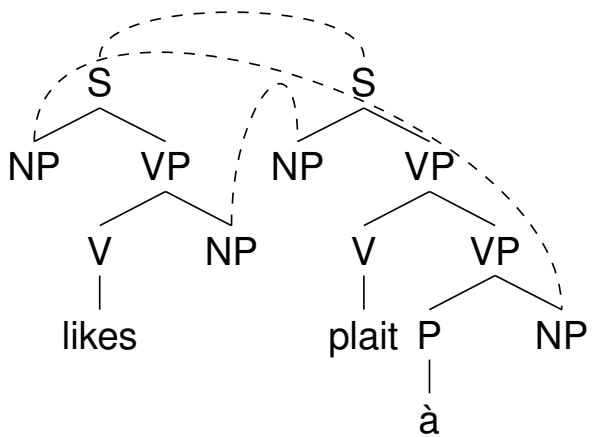
DOT: composition

The Tree-DOT composition operation ($\circ$):

- compose at the leftmost site on the fragment's source side
- compose at the target site *linked to* the leftmost source site

This ensures that:

- each derivation is unique
- translational equivalences encoded in the example base are respected



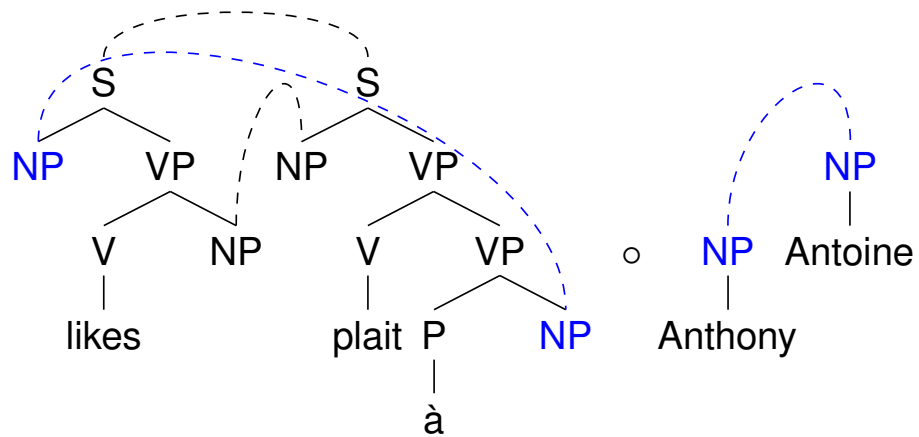
DOT: composition

The Tree-DOT composition operation (∘):

- compose at the leftmost site on the fragment's source side
- compose at the target site *linked to* the leftmost source site

This ensures that:

- each derivation is unique
- translational equivalences encoded in the example base are respected



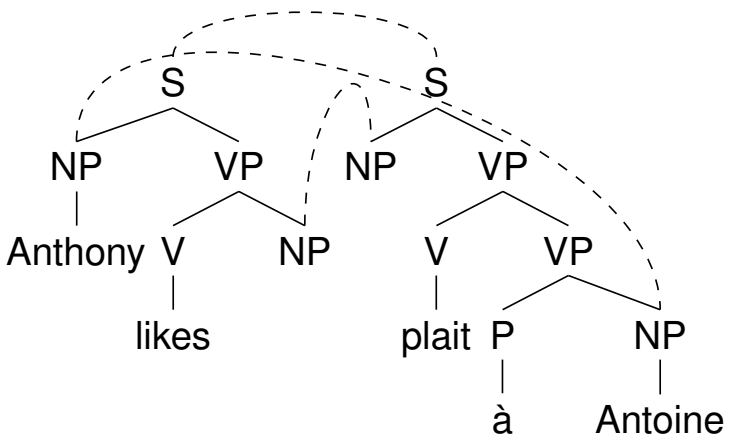
DOT: composition

The Tree-DOT composition operation (∘):

- compose at the leftmost site on the fragment's source side
- compose at the target site *linked to* the leftmost source site

This ensures that:

- each derivation is unique
- translational equivalences encoded in the example base are respected



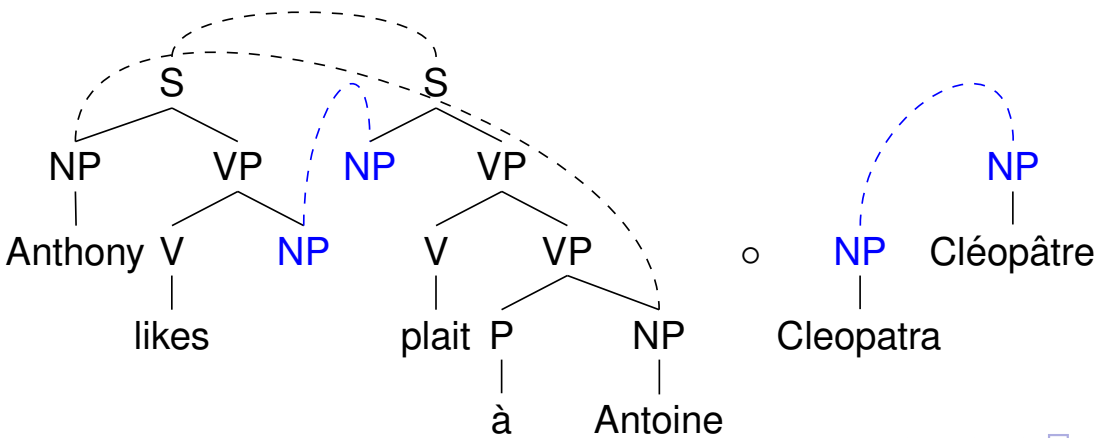
DOT: composition

The Tree-DOT composition operation (∘):

- compose at the leftmost site on the fragment's source side
- compose at the target site *linked to* the leftmost source site

This ensures that:

- each derivation is unique
- translational equivalences encoded in the example base are respected



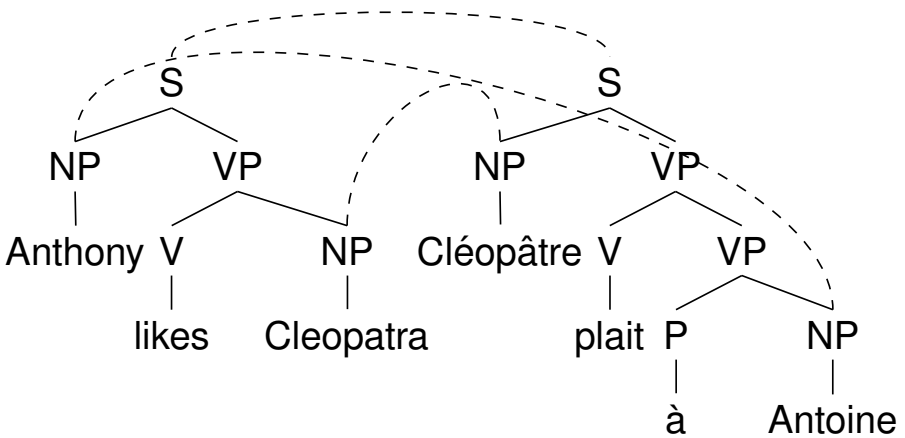
DOT: composition

The Tree-DOT composition operation (∘):

- compose at the leftmost site on the fragment's source side
- compose at the target site *linked to* the leftmost source site

This ensures that:

- each derivation is unique
- translational equivalences encoded in the example base are respected



DOT: probability model

DOT is a sum-of-products model

- probability of fragment $< s_x, t_x >$
- probability of derivation D_x
- probability of parse $< S_x, T_x >$
- probability of sentence pair s, t

$$\sum_{<S_x,T_x> \text{ yields } s,t} \quad \sum_{D_x \text{ yields } <S_x,T_x>} \quad \prod_{<s_x,t_x> \in D_x} \frac{|<s_x,t_x>|}{\sum_{root(s)=root(s_x) \wedge root(t)=root(t_x)} |<s,t>|}$$

DOT: probability model

DOT is a sum-of-products model

- probability of fragment $< s_x, t_x >$
- probability of derivation D_x
- probability of parse $< S_x, T_x >$
- probability of sentence pair s, t

$$\sum_{<S_x,T_x> \text{ yields } s,t} \quad \sum_{D_x \text{ yields } <S_x,T_x>} \quad \prod_{<s_x,t_x> \in D_x} \frac{|<s_x,t_x>|}{\sum_{root(s)=root(s_x) \wedge root(t)=root(t_x)} |<s,t>|}$$

DOT: probability model

DOT is a sum-of-products model

- probability of fragment $< s_x, t_x >$
- probability of derivation D_x
- probability of parse $< S_x, T_x >$
- probability of sentence pair s, t

$$\sum_{<S_x,T_x> \text{ yields } s,t} \quad \sum_{D_x \text{ yields } <S_x,T_x>} \quad \prod_{<s_x,t_x> \in D_x} \frac{|<s_x,t_x>|}{\sum_{root(s)=root(s_x) \wedge root(t)=root(t_x)} |<s,t>|}$$

DOT: probability model

DOT is a sum-of-products model

- probability of fragment $< s_x, t_x >$
- probability of derivation D_x
- probability of parse $< S_x, T_x >$
- probability of sentence pair s, t

$$\sum_{<S_x,T_x> \text{ yields } s,t} \quad \sum_{D_x \text{ yields } <S_x,T_x>} \quad \prod_{<s_x,t_x> \in D_x} \frac{|<s_x,t_x>|}{\sum_{root(s)=root(s_x) \wedge root(t)=root(t_x)} |<s,t>|}$$

DOT: probability model

DOT is a sum-of-products model

- probability of fragment $< s_x, t_x >$
- probability of derivation D_x
- probability of parse $< S_x, T_x >$
- probability of sentence pair s, t

$$\sum_{<S_x,T_x> \text{ yields } s,t} \quad \sum_{D_x \text{ yields } <S_x,T_x>} \quad \prod_{<s_x,t_x> \in D_x} \frac{|<s_x,t_x>|}{\sum_{root(s)=root(s_x) \wedge root(t)=root(t_x)} |<s,t>|}$$

Outline

Disambiguation Strategies

Strategies:

MPT: the **most probable sequence of target terminals** given the input string;

MPP: the sequence of target terminals read from the **most probable bilingual representation** for the input string;

MPD: the sequence of target terminals read from the **most probable derivation of a bilingual representation** for the input string;

SDER: the sequence of target terminals read from the **shortest derivation of a bilingual representation** for the input string.

Disambiguation Strategies

Strategies:

MPT: the **most probable sequence of target terminals** given the input string;

MPP: the sequence of target terminals read from the **most probable bilingual representation** for the input string;

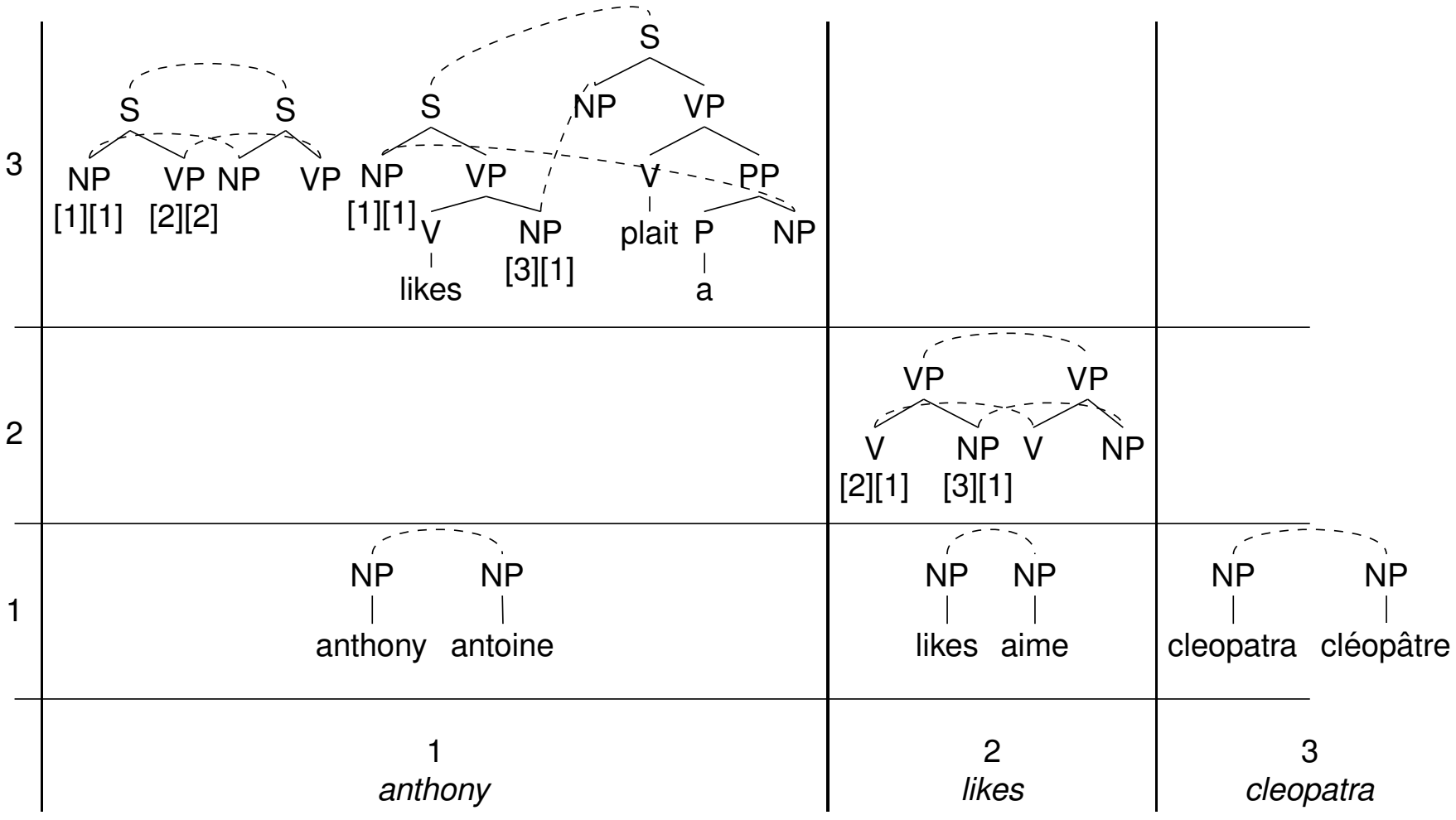
Monte Carlo Sampling

MPD: the sequence of target terminals read from the **most probable derivation of a bilingual representation** for the input string;

SDER: the sequence of target terminals read from the **shortest derivation of a bilingual representation** for the input string.

Viterbi

Translation Space



Monte Carlo Sampling [1/2]

Basic idea:

- rank the possible translations (MPT) or representations (MPP) according to how often each one occurs in a reduced random sample of the possible derivations
- the most frequently-occurring translation or representation corresponds to the most probable one according to the DOT model

Sampling a single derivation:

- select and compose fragments from the translation space top-down left-to-right until no open substitution sites remain
- fragments are selected such that if the DOP probability of fragment f_x is n times that of f_y, then f_x is n times more likely to be chosen during random selection than f_y

Monte Carlo Sampling [2/2]

Computing a set of sampled derivations:

- must determine the minimum number of samples needed to be certain that the most frequent solution in the sampled set is the DOT most probable one
- BKS: sample until we fulfil a stopping condition which is predefined but recalculated each time a new sample is added to the set

The decision to stop sampling is based on:

- how closely matched, in terms of frequency of occurrence, the translations in the sample set are
- how many of the possible translations for the given input string are present in the set of sampled translations
- how certain we wish to be that the most frequent translation in the sample set is in fact the most probable translation according to the DOT model

Viterbi

Computing the MPD using Viterbi:

- prune sub-derivations with low probabilities from the translation space, i.e.
- at each chart position, retain only the most probable sub-derivation for each possible syntactic category

Computing the SDER using Viterbi:

- assign each fragment equal probability $\frac{1}{p}$
- the probability of a derivation using n fragments is $(\frac{1}{p})^n$
- the smaller n will give a larger derivation probability, thus the shortest derivation is computed as the most probable one

Outline

Experiments [1/2]

Data:

- English-French section of the HomeCentre corpus
- 810 parsed, sub-sententially aligned translation pairs
- split into 12 training/test sets:
  - 80 test sentence/reference pairs and 730 training tree pairs
  - 6 splits English-to-French, 6 splits French-to-English
  - randomly split such that all test words occur in the training set (no OOV items)

Experiments [2/2]

Runs:

- MPT, MPP, MPD, SDER
- link depths 1, ≤ 2, ≤ 3, ≤ 4

	depth=1	depth$\leq$2	depth$\leq$3	depth$\leq$4
EN-to-FR:	6,140	29,081	148,165	1,956,786
FR-to-EN:	6,197	29,355	150,460	2,012,632

Evaluation metrics:

Exact match, Bleu and F-score, averaged over splits

Results – Accuracy, English-to-French

The DOP Hypothesis:

- as fragment depth increases, accuracy increases

Disambiguation strategies:

- All metrics and depths: either MPD or SDER is preferred (exception: MPP d2)
- MPT: does not achieve highest accuracy at any depth for any metric
- Overall: highest performance is at depth 4 using MPD or SDER

		$=1$	≤ 2	≤ 3	≤ 4
BLEU	MPT	0.4479	0.5034	0.5277	0.5343
	MPP	0.4507	0.4946	0.5192	0.5216
	MPD	0.4572	0.5069	0.5269	0.5386
	SDER	0.4168	0.5080	0.5314	0.5386
F-score	MPT	0.6712	0.7035	0.7179	0.7222
	MPP	0.6733	0.6990	0.7135	0.7149
	MPD	0.6793	0.7083	0.7213	0.7257
	SDER	0.6513	0.7074	0.7204	0.7254
Exact match	MPT	30.21	37.92	40.00	41.25
	MPP	30.62	37.50	38.96	40.00
	MPD	30.42	37.08	39.17	41.04
	SDER	25.62	38.12	41.46	42.29

Results – Accuracy, English-to-French

The DOP Hypothesis:

- as fragment depth increases, accuracy increases

Disambiguation strategies:

- All metrics and depths: either MPD or SDER is preferred (exception: MPP d2)
- MPT: does not achieve highest accuracy at any depth for any metric
- Overall: highest performance is at depth 4 using MPD or SDER

		$=1$	≤ 2	≤ 3	≤ 4
BLEU	MPT	0.4479	0.5034	0.5277	0.5343
	MPP	0.4507	0.4946	0.5192	0.5216
	MPD	0.4572	0.5069	0.5269	0.5386
	SDER	0.4168	0.5080	0.5314	0.5386
F-score	MPT	0.6712	0.7035	0.7179	0.7222
	MPP	0.6733	0.6990	0.7135	0.7149
	MPD	0.6793	0.7083	0.7213	0.7257
	SDER	0.6513	0.7074	0.7204	0.7254
Exact match	MPT	30.21	37.92	40.00	41.25
	MPP	30.62	37.50	38.96	40.00
	MPD	30.42	37.08	39.17	41.04
	SDER	25.62	38.12	41.46	42.29

Results – Accuracy, English-to-French

The DOP Hypothesis:

- as fragment depth increases, accuracy increases

Disambiguation strategies:

- All metrics and depths: either MPD or SDER is preferred (exception: MPP d2)
- MPT: does not achieve highest accuracy at any depth for any metric
- Overall: highest performance is at depth 4 using MPD or SDER

		$=1$	≤ 2	≤ 3	≤ 4
BLEU	MPT	0.4479	0.5034	0.5277	0.5343
	MPP	0.4507	0.4946	0.5192	0.5216
	MPD	0.4572	0.5069	0.5269	0.5386
	SDER	0.4168	0.5080	0.5314	0.5386
F-score	MPT	0.6712	0.7035	0.7179	0.7222
	MPP	0.6733	0.6990	0.7135	0.7149
	MPD	0.6793	0.7083	0.7213	0.7257
	SDER	0.6513	0.7074	0.7204	0.7254
Exact match	MPT	30.21	37.92	40.00	41.25
	MPP	30.62	37.50	38.96	40.00
	MPD	30.42	37.08	39.17	41.04
	SDER	25.62	38.12	41.46	42.29

Results – Accuracy, French-to-English

The DOP Hypothesis:

- as fragment depth increases, accuracy increases

Disambiguation strategies:

- Bleu and F-score: MPT scores best (exception: SDER Bleu d3)
- Exact match: little consistency – SDER at depths 3 and 4

		$=1$	≤ 2	≤ 3	≤ 4
BLEU	MPT	0.4990	0.5513	0.5447	0.5494
	MPP	0.4915	0.5406	0.5454	0.5449
	MPD	0.4946	0.5396	0.5436	0.5434
	SDER	0.4316	0.5318	0.5465	0.5488
F-score	MPT	0.7177	0.7463	0.7443	0.7463
	MPP	0.7098	0.7407	0.7423	0.7427
	MPD	0.7119	0.7376	0.7386	0.7396
	SDER	0.6832	0.7343	0.7401	0.7421
Exact match	MPT	43.75	49.17	48.75	49.38
	MPP	44.38	50.00	49.38	50.21
	MPD	44.79	49.38	49.79	50.21
	SDER	36.46	48.54	50.00	50.42

Results – Accuracy, French-to-English

The DOP Hypothesis:

- as fragment depth increases, accuracy increases

Disambiguation strategies:

- Bleu and F-score: MPT scores best (exception: SDER Bleu d3)
- Exact match: little consistency – SDER at depths 3 and 4

		$=1$	≤ 2	≤ 3	≤ 4
BLEU	MPT	0.4990	0.5513	0.5447	0.5494
	MPP	0.4915	0.5406	0.5454	0.5449
	MPD	0.4946	0.5396	0.5436	0.5434
	SDER	0.4316	0.5318	0.5465	0.5488
F-score	MPT	0.7177	0.7463	0.7443	0.7463
	MPP	0.7098	0.7407	0.7423	0.7427
	MPD	0.7119	0.7376	0.7386	0.7396
	SDER	0.6832	0.7343	0.7401	0.7421
Exact match	MPT	43.75	49.17	48.75	49.38
	MPP	44.38	50.00	49.38	50.21
	MPD	44.79	49.38	49.79	50.21
	SDER	36.46	48.54	50.00	50.42

Results – Accuracy, French-to-English

The DOP Hypothesis:

- as fragment depth increases, accuracy increases

Disambiguation strategies:

- Bleu and F-score: MPT scores best (exception: SDER Bleu d3)
- Exact match: little consistency – SDER at depths 3 and 4

		$=1$	≤ 2	≤ 3	≤ 4
BLEU	MPT	0.4990	0.5513	0.5447	0.5494
	MPP	0.4915	0.5406	0.5454	0.5449
	MPD	0.4946	0.5396	0.5436	0.5434
	SDER	0.4316	0.5318	0.5465	0.5488
F-score	MPT	0.7177	0.7463	0.7443	0.7463
	MPP	0.7098	0.7407	0.7423	0.7427
	MPD	0.7119	0.7376	0.7386	0.7396
	SDER	0.6832	0.7343	0.7401	0.7421
Exact match	MPT	43.75	49.17	48.75	49.38
	MPP	44.38	50.00	49.38	50.21
	MPD	44.79	49.38	49.79	50.21
	SDER	36.46	48.54	50.00	50.42

Results – Efficiency

Observations:

- time increases as depth increases
  - extra time is spent building the translation space rather than disambiguating
  - translating from French takes longer because avg. sentence length is longer
- En-to-Fr: SDER = MPD $<$ MPP $<$ MPT
- Fr-to-En: MPT $<$ SDER = MPD $<$ MPP
- Overall: ranking with Monte Carlo does not take longer than ranking with Viterbi for this dataset

ENGLISH-TO-FRENCH					FRENCH-TO-ENGLISH				
	CPU seconds/sentence					CPU seconds/sentence			
	MPT	MPP	MPD	SDER		MPT	MPP	MPD	SDER
1	1.39	1.33	0.29	0.30	1	0.72	3.73	3.12	3.13
2	2.06	1.55	0.57	0.58	2	1.16	3.85	3.53	3.58
3	3.05	2.28	1.40	1.41	3	2.32	4.96	4.62	4.64
4	12.8	11.9	11.3	11.1	4	18.9	21.5	21.1	20.8

Current Work

- move to larger datasets and different language pairs
  - improve (speed) algorithm to build translation space
  - improve automatic sub-structural alignment
- test for statistical significance
- compare against phrase-based SMT

- try probabilistic SDER, as proposed for DOP
- improve parameter estimation using methods proposed for DOP
- incorporate LFG representations