Comparing Constituency and Dependency Representations for SMT Phrase Extraction

Sylwia Ozdowska
(joint work with Mary Hearne and John Tinsley)

23rd July 2008

Phrase-pair Extraction and Syntax

- Standard technique to induce translation models (string-based phrase-pairs) not syntax-aware (Koehn et al. 03)

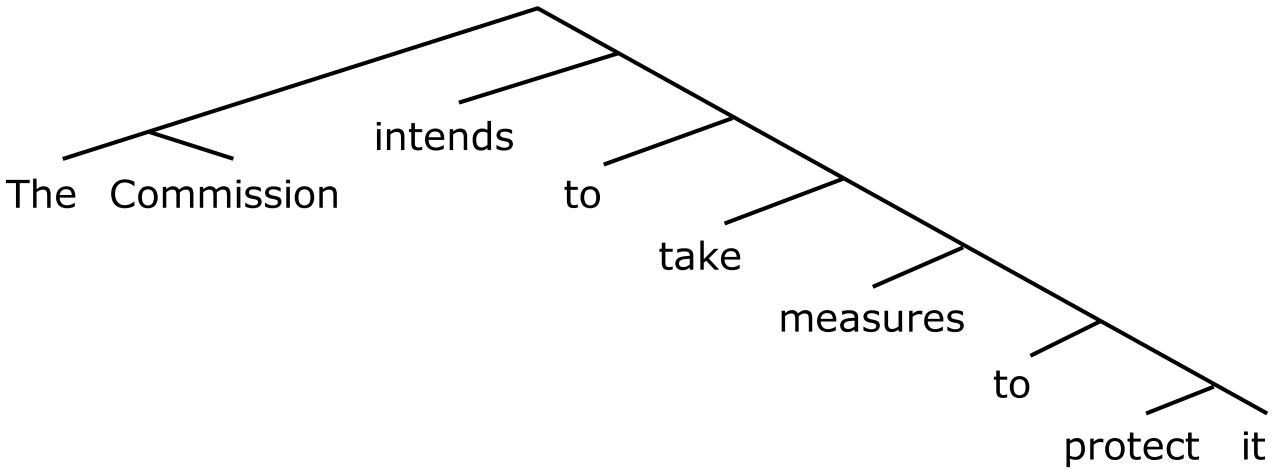
  Q: What about adding syntax into the phrase extraction process?

- Combining constituency-based phrase pairs with string-based phrase pairs improves translation quality (Tinsley et al. 07)

  Q: What about using dependency-based phrase pairs?
  - What is the value of replacing and/or combining string-based methods with syntax-based methods for PB-SMT?
  - What are the relative merits of using constituency-annotated *vs.* dependency-annotated training data?

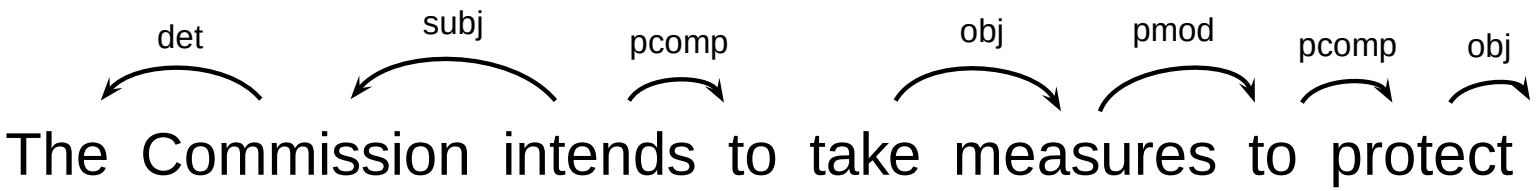
Constituency parses

- Constituency parses are context-free phrase-structure trees
- They make explicit syntactic constituents such as noun phrases (*eg. the commission*), verb phrases (*eg. intends to take. . .*), prepositional phrases (*eg. to protect it*), etc.



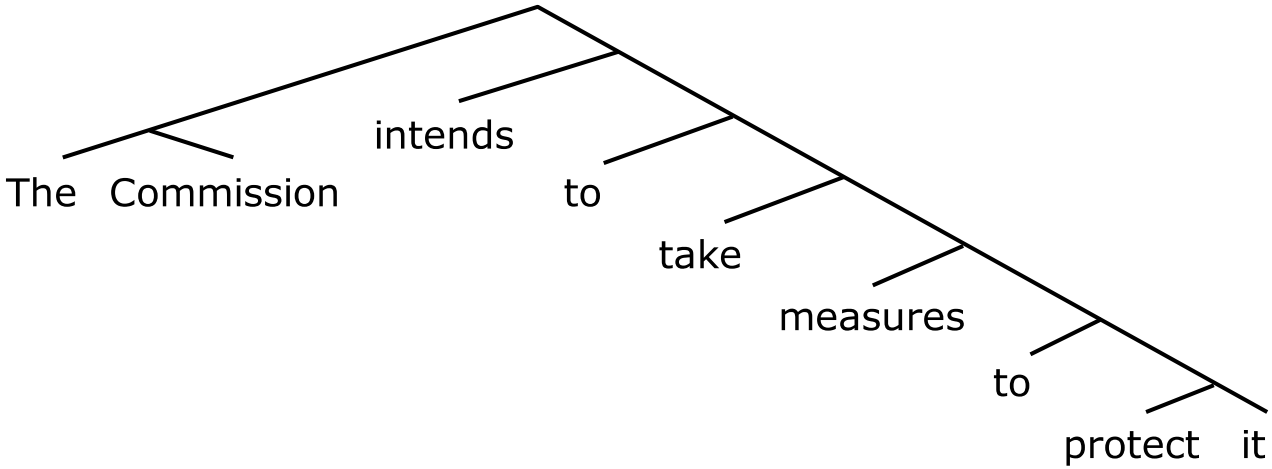
Dependency parses

- Dependency parses make explicit the relationships between the words in terms of heads and dependents and possibly the nature of the relationship, *ie.* subject (*eg. Commission* depends on *intends* via the subject relation), object (*eg. measures* depends on *take* via the object relation), etc.

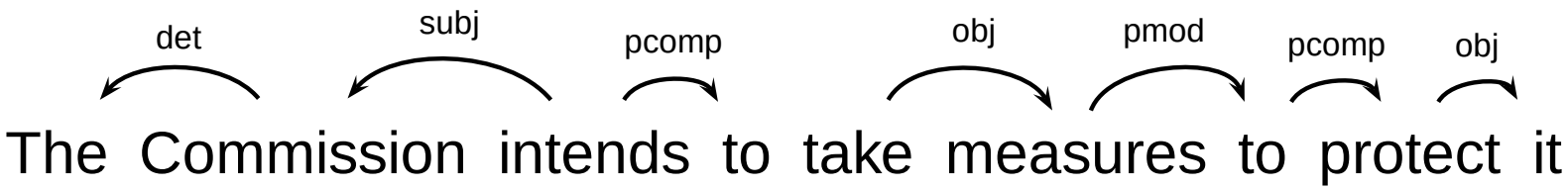


Recap: constituency parses and dependency parses

- Constituency parses



- Dependency parses

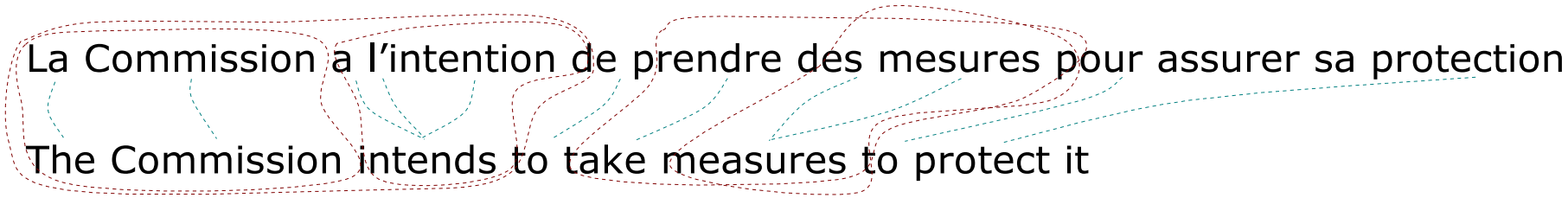


Parsing

- Corpus
  - JOC English–French parallel corpus (Chiao et al. 06)
  - EU Parliamentary Proceedings (1993)
  - 8,759 aligned sentences (7722 training pairs and 1000 test pairs)
- Constituency annotation: Bikel's statistical parser (Bikel 02)
  - English: training on the Penn II Treebank (Marcus et al. 96)
  - French: training on the Modified French Treebank (Schulter & Genabith 07)
- Dependency annotation: Syntex parsers (Bourigault et al. 05)

String-based alignment and phrase extraction (STR)

- Giza++ (Och & Ney 03), Moses (Koehn et al. 07)
- Extraction of word and phrase-pairs consistent with the word alignment

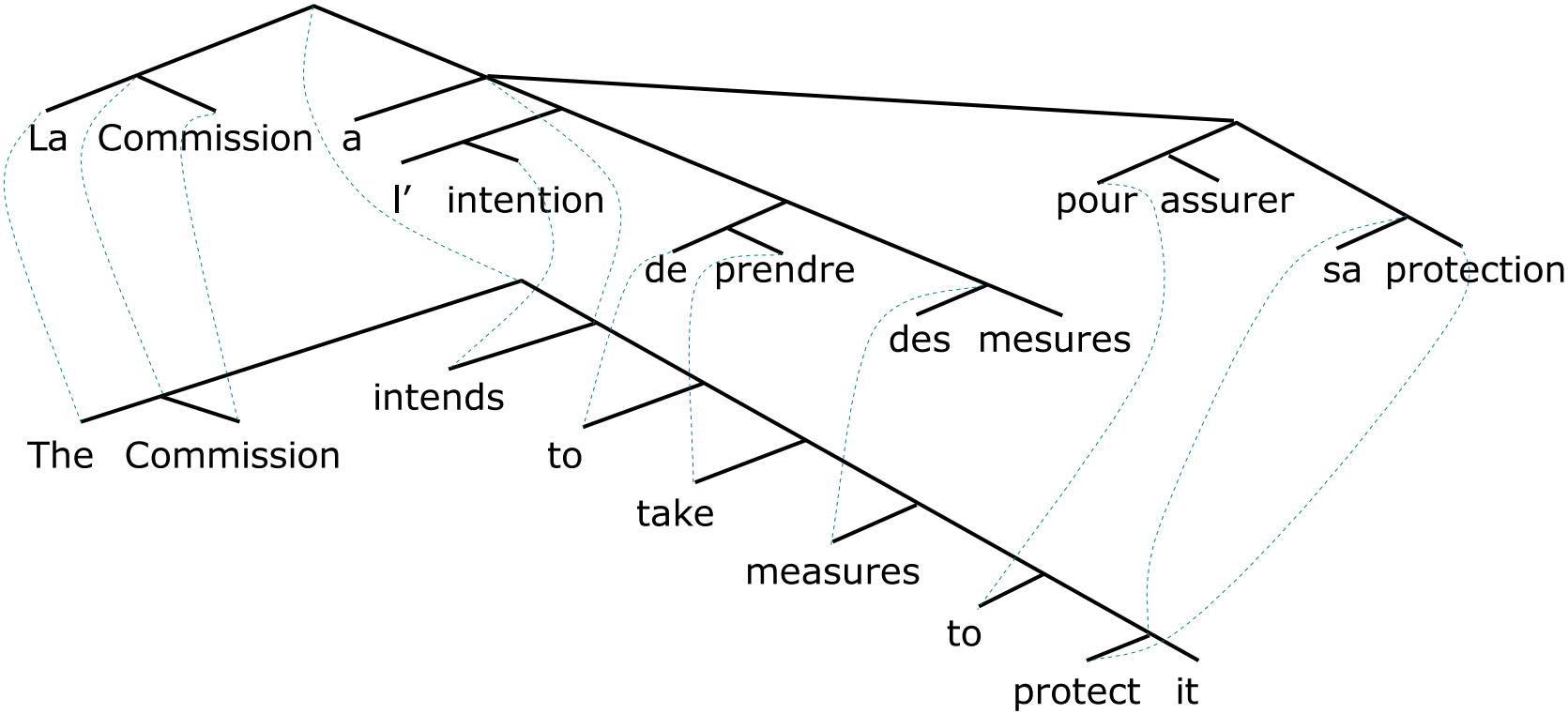


a l'intention $\Leftrightarrow$ intends
a l'intention de $\Leftrightarrow$ intends to
commission a l'intention de $\Leftrightarrow$ commission intends to
la commission a l'intention de $\Leftrightarrow$ the commission intends to

Syntax-based alignment and phrase-pair extraction

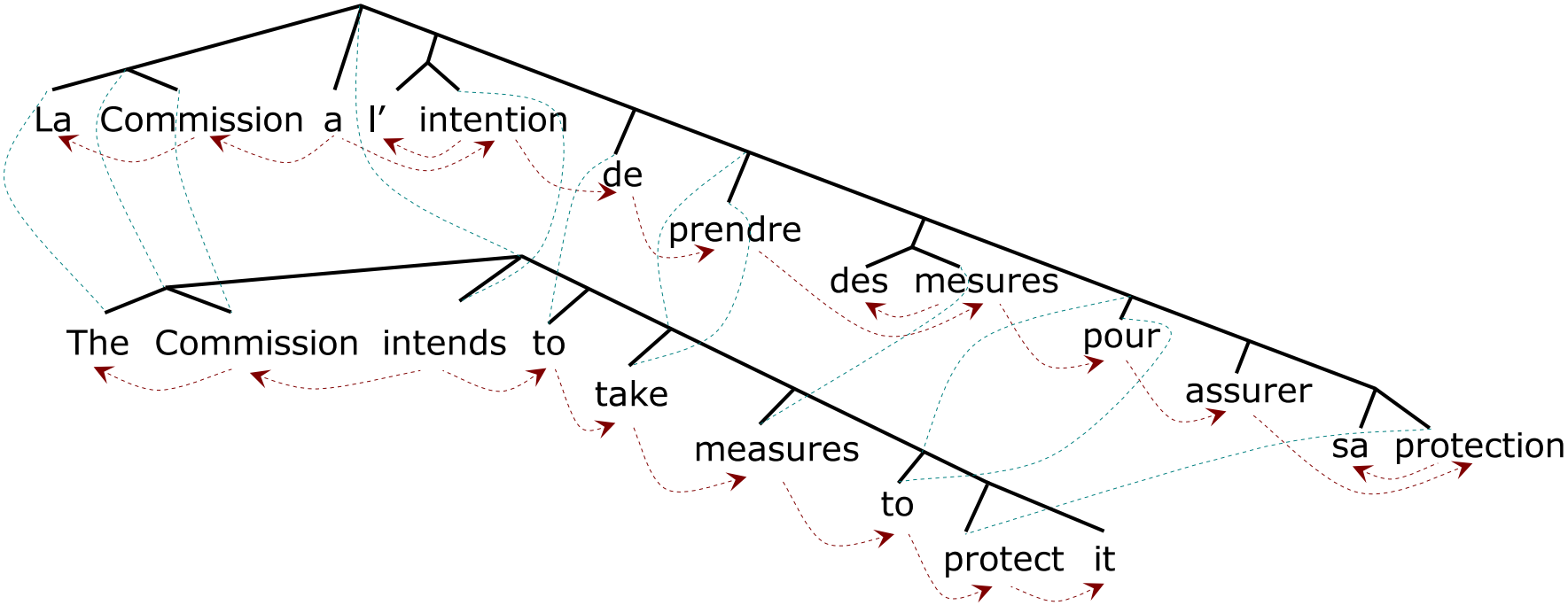
- TreeAligner (Tinsley et al. 07)
- Extraction of all string pairs dominated by linked constituents

Constituency-based (CON)



de prendre des mesures pour assurer sa protection ⇔ to take measures to protect it
pour assurer sa protection ⇔ to protect it
des mesures ⇔ measures

Dependency-based (DEP)



a l'intention de prendre des mesures pour assurer sa protection ⇔ intends to take measures to protect it

sa protection ⇔ protect it

mesures ⇔ measures

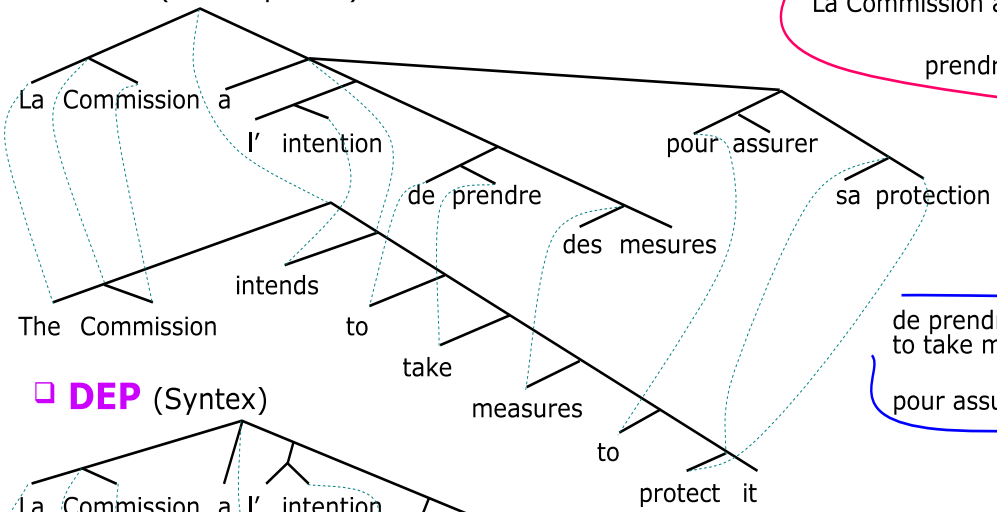
Recap: STR, CON and DEP

- **STR**

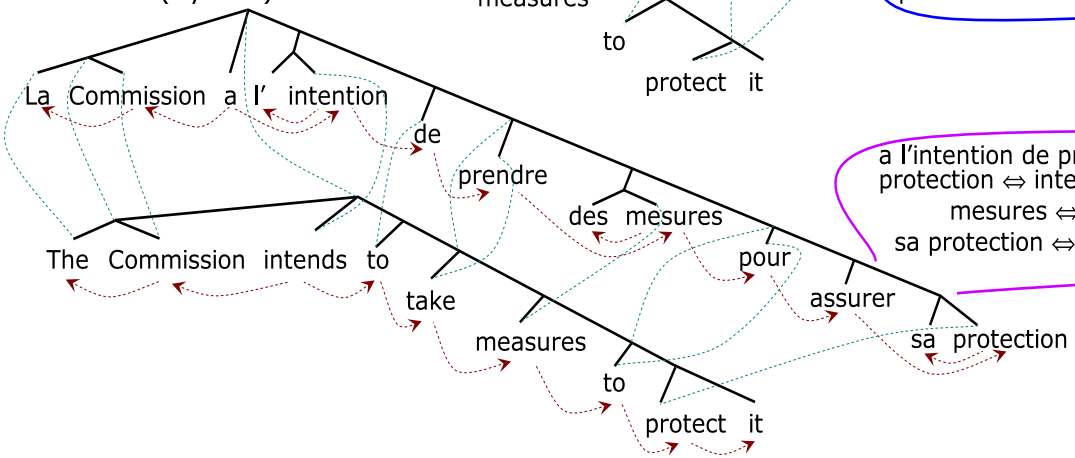
La Commission a l'intention de prendre des mesures pour assurer sa protection

The Commission intends to take measures to protect it

- **CON** (Bikel's parser)



- **DEP** (Syntex)



a l'intention ⇔ intends
a l'intention de ⇔ intends to
Commission a l'intention de ⇔ Commission intends to
La Commission a l'intention de ⇔ The Commission intends to
des mesures ⇔ measures
prendre des mesures ⇔ take measures
... ⇔ ...

de prendre des mesures pour assurer sa protection ⇔ to take measures to protect it
des mesures ⇔ measures
pour assurer sa protection ⇔ to protect it

a l'intention de prendre des mesures pour assurer sa protection ⇔ intends to take measures to protect it
mesures ⇔ measures
sa protection ⇔ protect it

la ⇔ the
commission ⇔ commission
la commission ⇔ the commissio
intention ⇔ intends
de ⇔ to
prendre ⇔ take
pour ⇔ to
protection ⇔ protect

Translation

- 1000 sentences
- Moses decoder (Koehn et al. 07)
- Different phrase-table configurations
  - STR|CON|DEP: each phrase table used individually
  - STR+(CON|DEP): combination of STR with either CON or DEP
  - STR+CON+DEP: combination of all phrase-pairs

Results

	BLEU	NIST	METEOR
STR	30.35	62.62	64.32
CON	29.82	62.96	63.53
DEP	29.80	63.07	64.00
STR+CON	31.88	65.04	65.59
STR+DEP	31.97	65.07	65.70
STR+CON+DEP	31.90	65.10	65.57

- Conflicting results across evaluation metrics
- Replacing STR with CON|DEP: translation quality$--$
- Combining STR with CON|DEP: translation quality++
- Translation quality CON $<$ Translation quality DEP

Analysis

- Phrase-pair length
  - shorter phrase-pairs $>$ greater impact on transaltion quality
  - unique phrase-pair average length: 9.98 for CON *vs.* 5.67 for DEP
- Phrase-pair coverage
  - higher alignment coverage $>$ lower phrase-pair quality
  - linked phrase-pairs:
    66,601 En and 67,280 Fr for CON
    64,904 En and 64,135 Fr for DEP
- Parsers' performance
  - different monolingual parses $>$ differing translation accuracies
  - parsers' f-scores:
    90% En and 80% Fr for CON
    82% En and 89% Fr for DEP
- Generated phrases 52% En and 46.6% Fr unique to CON
  38.9% En and 42.7% Fr unique to DEP

This is the end. . .

- Conclusions
  - Translation quality improves when combining STR with either CON or DEP
  - General trend towards preferring dependency-based phrase-pairs over constituency-based phrase-pairs
- Future work
  - Gain further insights into the reasons why translation quality varies
  - Analyse the relative impact of the different types of constituents for which phrase-pairs are extracted
  - Scale up the experiments