Parallel Treebanks in MT & MaTrEx: The DCU MT System

John Tinsley

NCLT/CNGL Workshop

Talk Overview

PhD Work

- What I've done and why
- What I intend to do

MaTrEx

- Overview and functional description
- Resources
- Maintenance and continuous development

Motivation

- Output of the aligner
- State of the art MT

What is the SOTA?

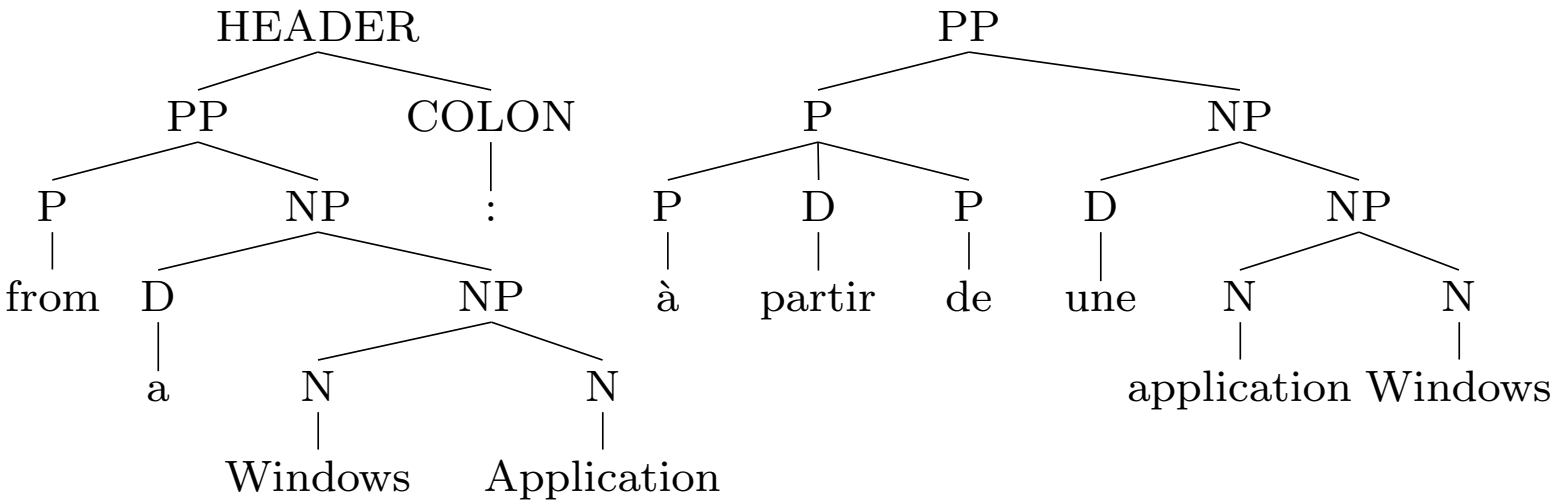
- Phrase-based statistical MT
- Parallel corpus + statistical word alignment + heuristics = translation model
- No linguistic motivation

Methodology

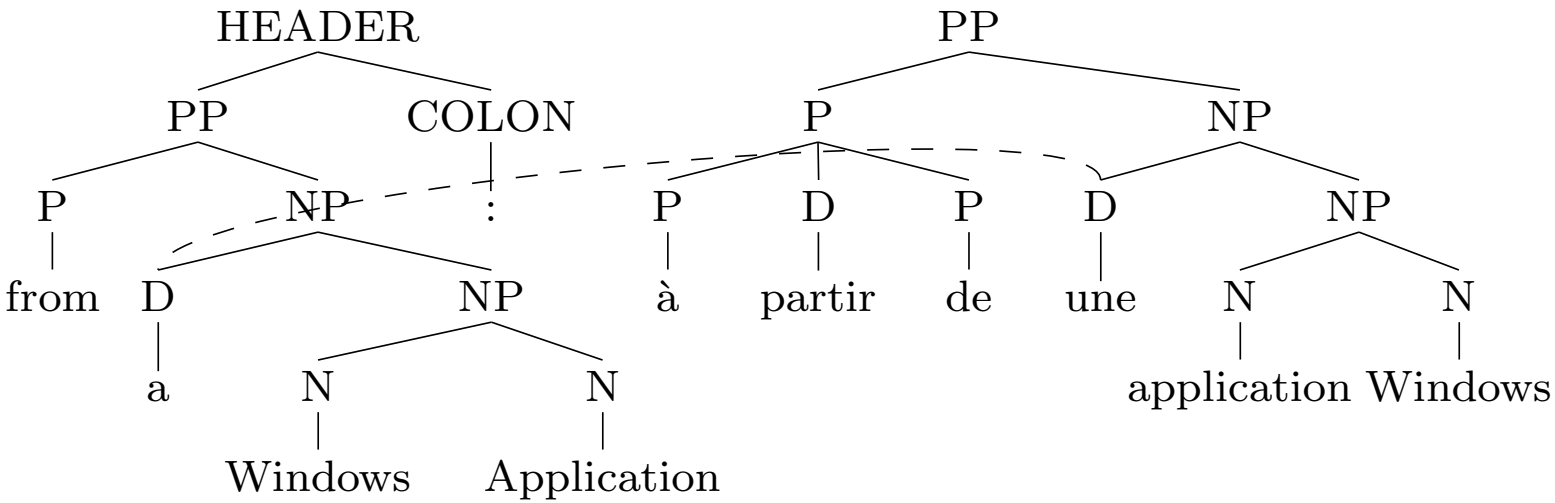
How do we use the treebank?

- Extract syntax-driven phrase pairs based on alignments
- Add them to the translation model of the SMT system
- Estimate new probabilities

Treebank Phrase Extraction

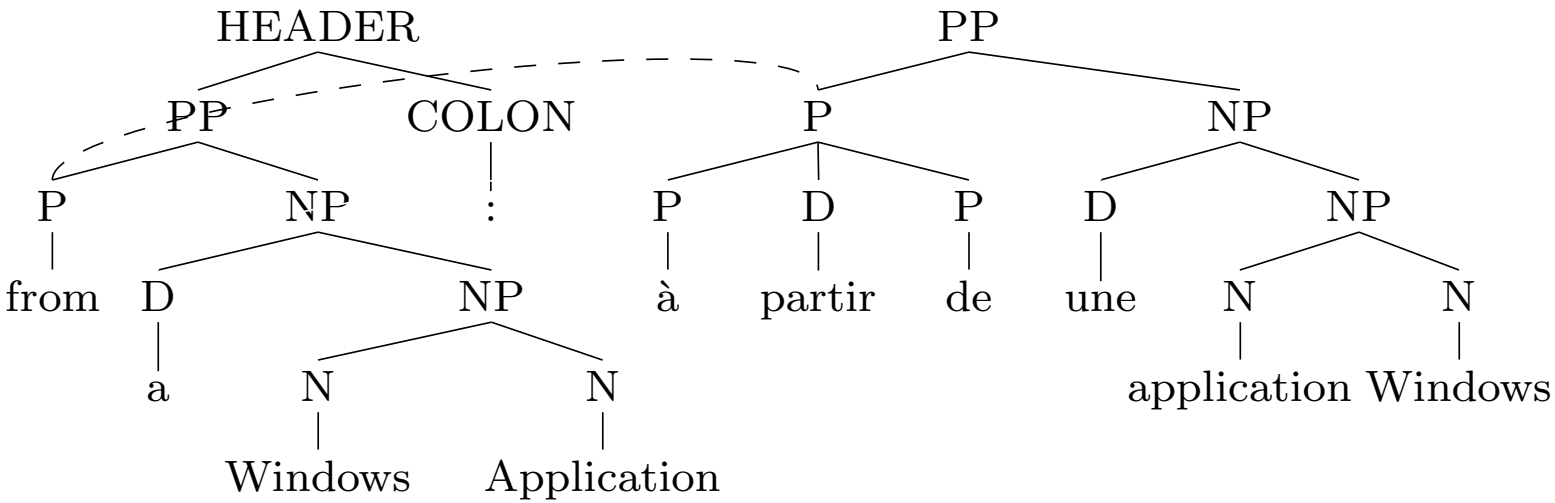


Treebank Phrase Extraction



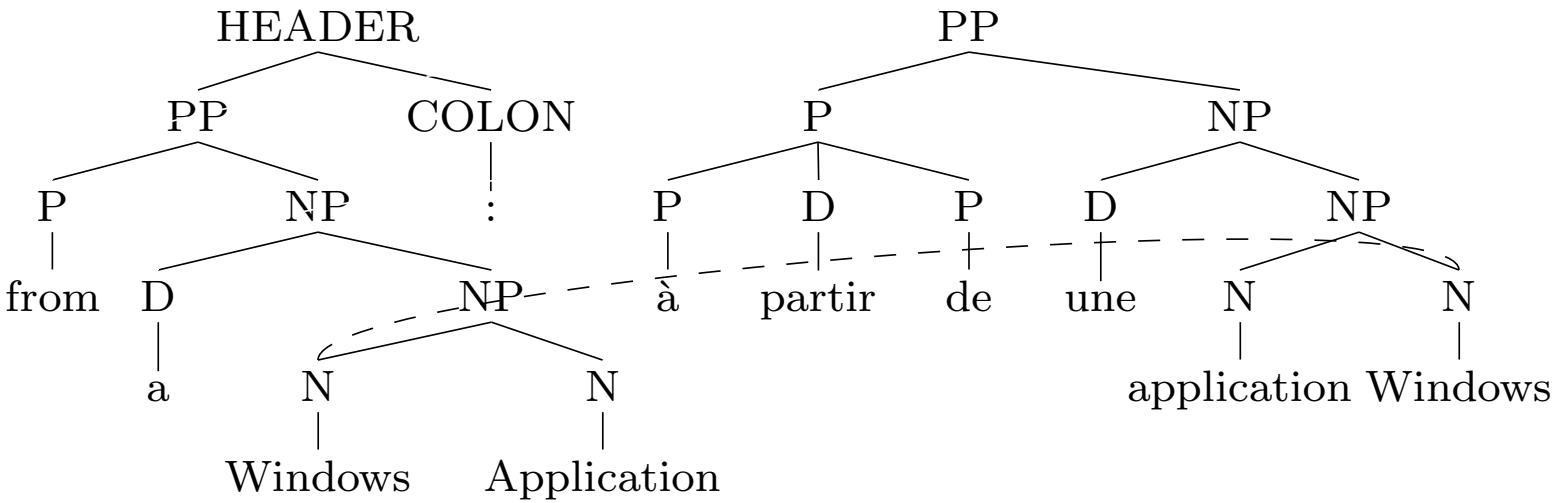
a ⇔ une

Treebank Phrase Extraction



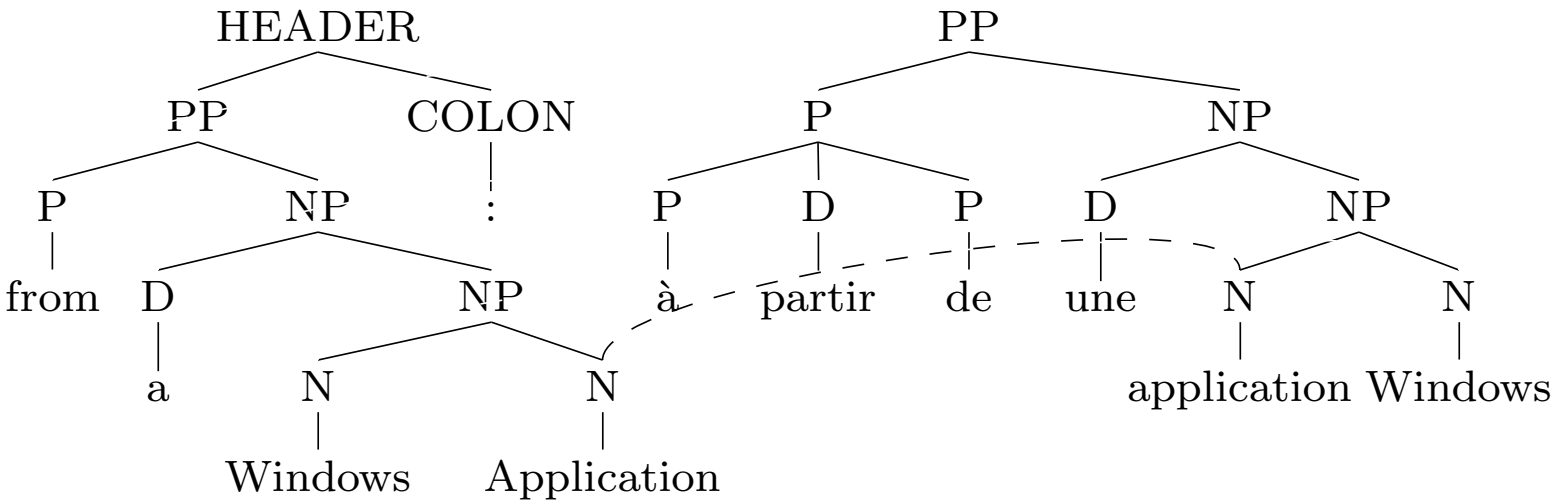
a	⇔	une
from	⇔	à partir de

Treebank Phrase Extraction



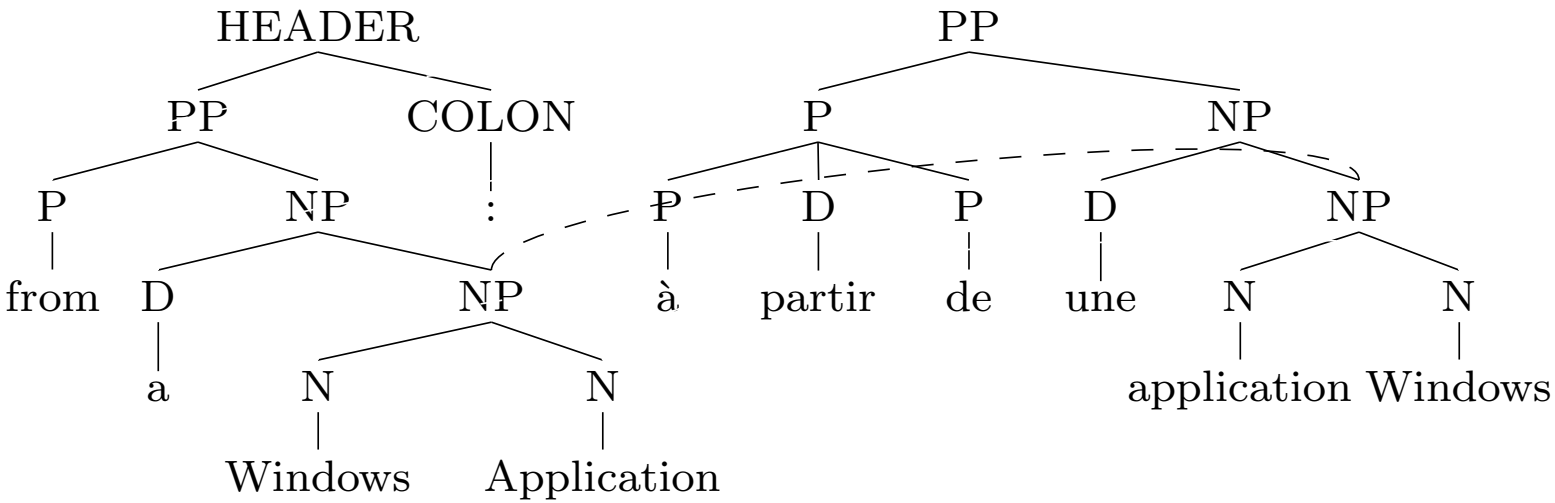
a	⇔	une
from	⇔	à partir de
Windows	⇔	Windows

Treebank Phrase Extraction



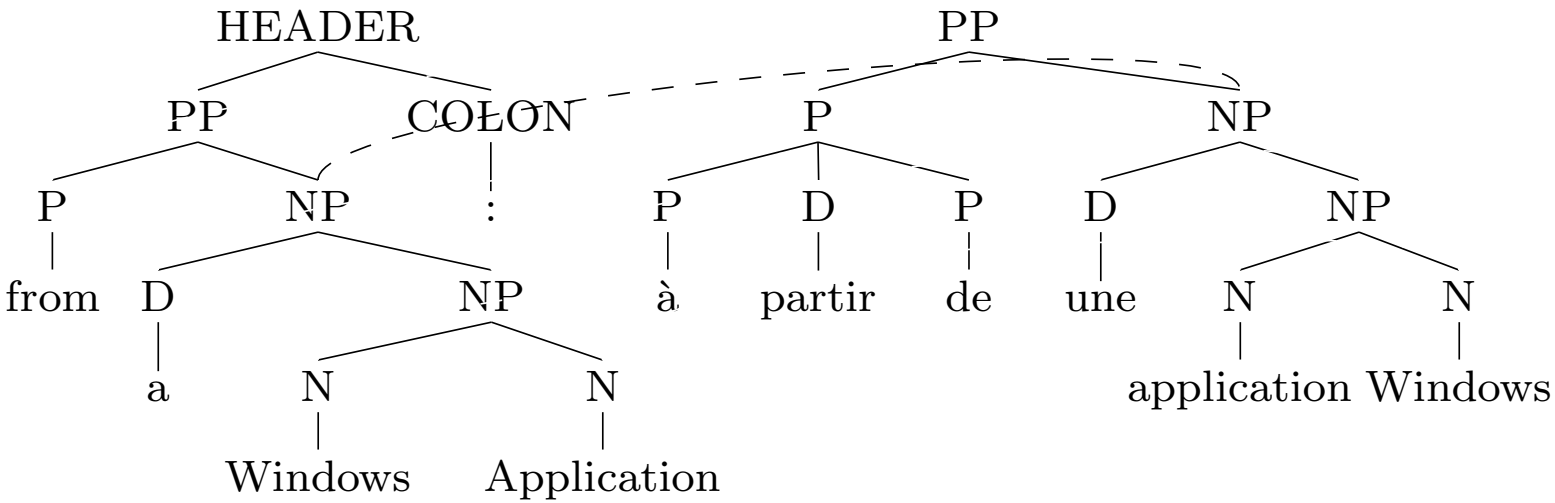
a	⇔	une
from	⇔	à partir de
Windows	⇔	Windows
Application	⇔	application

Treebank Phrase Extraction



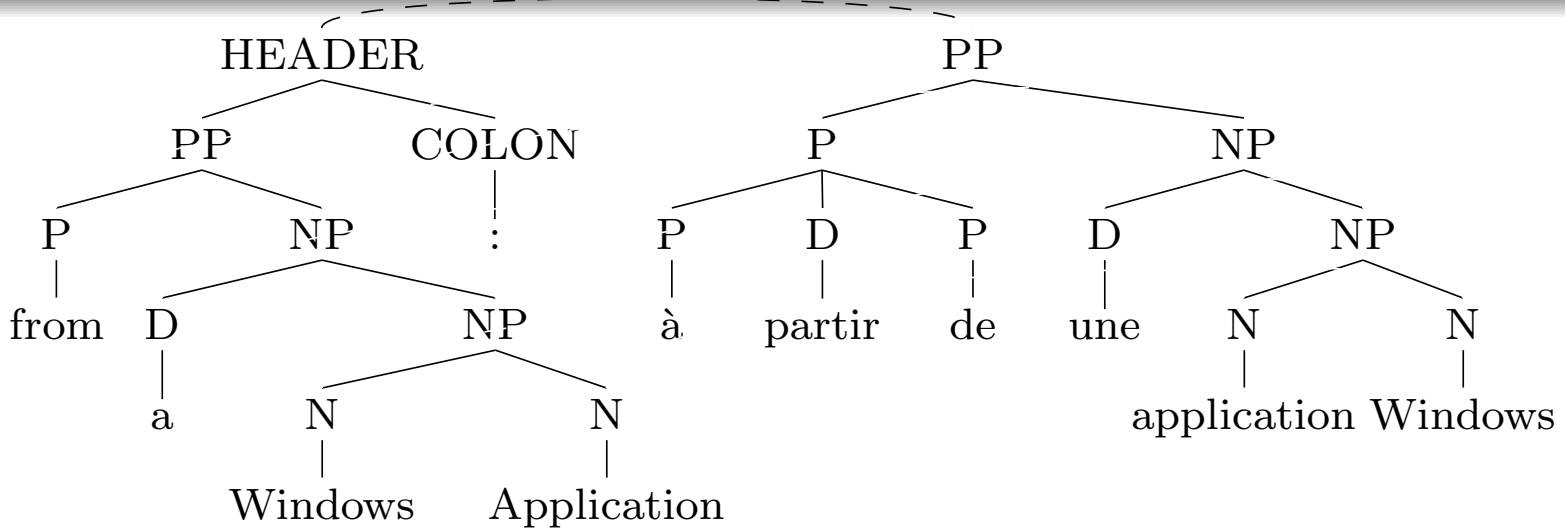
a	⇔	une
from	⇔	à partir de
Windows	⇔	Windows
Application	⇔	application
Windows Application	⇔	application Windows

Treebank Phrase Extraction



a	⇔	une
from	⇔	à partir de
Windows	⇔	Windows
Application	⇔	application
Windows Application	⇔	application Windows
a Windows Application	⇔	une application Windows

Treebank Phrase Extraction



a	⇔	une
from	⇔	à partir de
Windows	⇔	Windows
Application	⇔	application
Windows Application	⇔	application Windows
a Windows Application	⇔	une application Windows
from a Windows Application ;	⇔	à partir de une application Windows

Methodology

How do we use the treebank?

- Extract syntax-driven phrase pairs based on alignments
- Add them to the translation model of the SMT system
- Estimate new probabilities

Effect on translation model

- Increased coverage
- "Bonus" to relevant SMT phrase pairs

Outcome

How has this method performed?

- It *almost* always improves translation accuracy
- Effectiveness varies under different conditions

Outcome

How has this method performed?

- It *almost* always improves translation accuracy
- Effectiveness varies under different conditions

Language Pairs

English–Spanish √	Spanish–English √
English–French √	French–English √
English–German √	German–English √
English–Chinese ×	Chinese–English √ & ×
Chinese–Spanish √	

Outstanding Issues

What's left to do in terms of this work?

- Weighting of phrases (counting, lexical weights, multiple models)
- Reordering model for treebank phrases
- ...

So what now?

- So much more information in parallel treebanks
- Apply to more syntax-aware MT

How?

- Hierarchical SMT has shown improvements
- Allows for generalisation in phrase pairs
- Again unmotivated

Generalised Phrase Pairs (templates)

Training Sentence Pair

The team is good $\Leftrightarrow$ El equipo es bueno

Generalised Phrase Pairs (templates)

Training Sentence Pair

The team is good ⇔ El equipo es bueno

Extracted Phrase Pair

The team ⇔ El equipo

Generalised Phrase Pairs (templates)

Training Sentence Pair

The team is good ⇔ El equipo es bueno

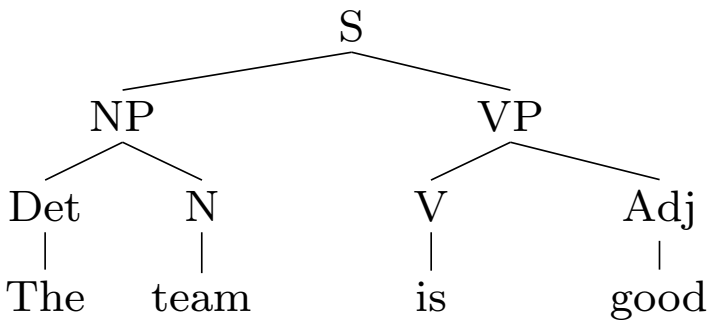
Extracted Phrase Pair

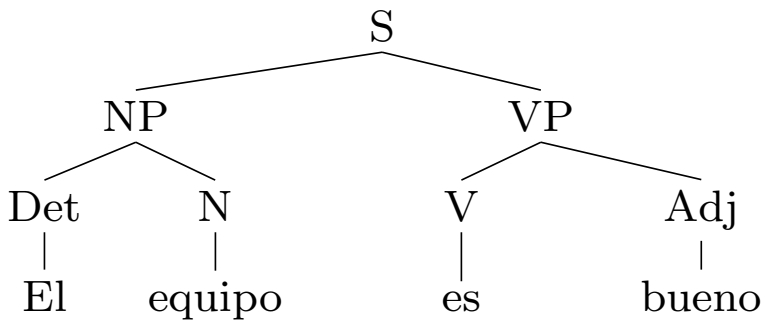
The team ⇔ El equipo

Generalised Template

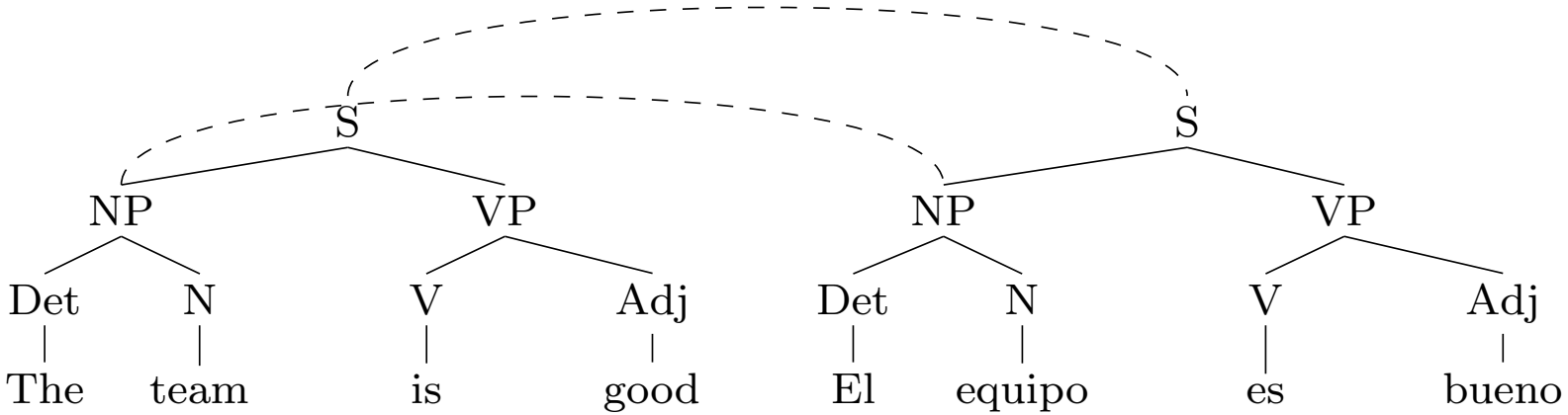
X is good ⇔ X es bueno

Generalised Phrase Pairs (templates)





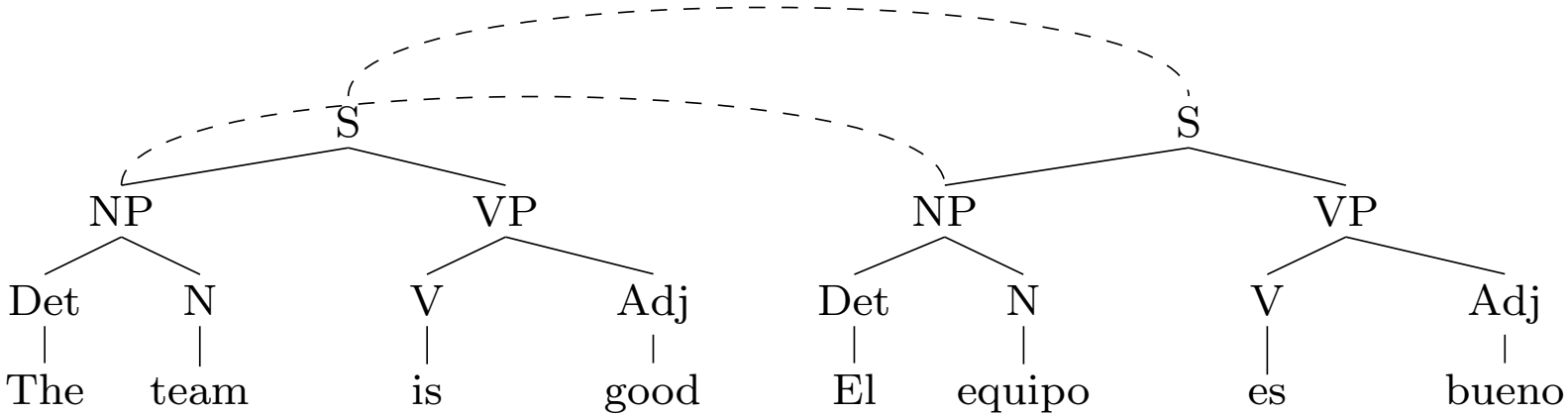
Generalised Phrase Pairs (templates)



The $team_{(NP)}$ $\Leftrightarrow$ El $equipo_{(NP)}$

The team is $good_{(S)}$ $\Leftrightarrow$ El equipo es $bueno_{(S)}$

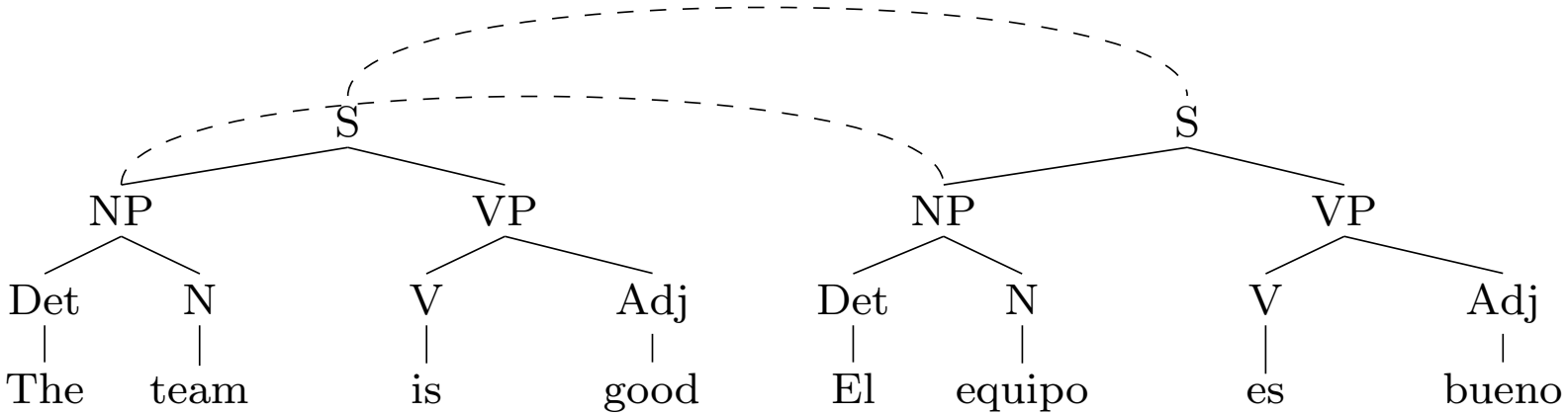
Generalised Phrase Pairs (templates)



$$\text{The } team_{(NP)} \Leftrightarrow \text{El } equipo_{(NP)}$$
$$\text{The team is } good_{(S)} \Leftrightarrow \text{El equipo es } bueno_{(S)}$$

$$\langle NP \rangle \text{ is good} \Leftrightarrow \langle NP \rangle \text{ es bueno}$$

Generalised Phrase Pairs (templates)



$$\text{The } team_{(NP)} \Leftrightarrow \text{El } equipo_{(NP)}$$
$$\text{The team is } good_{(S)} \Leftrightarrow \text{El equipo es } bueno_{(S)}$$

$$\langle NP \rangle \text{ is good} \Leftrightarrow \langle NP \rangle \text{ es bueno}$$

- We've constrained the generalisation

Application of Templates

How do we use these templates?

- Phrase-based decoders don't allow for templates
- Chiang's hierarchical decoder
- CMU's Stat-XFer system

Questions on this part?

MaTrEx: The DCU MT System

System Overview

What is MaTrEx?

- **Ma**chine **Tr**anslation using **Ex**amples
- Modular data-driven MT system
- Core = Makefile
- Wrapper around extendible and reimplementable modules

System Overview

What is MaTrEx?

- **Ma**chine **Tr**anslation using **Ex**amples
- Modular data-driven MT system
- Core = Makefile
- Wrapper around extendible and reimplementable modules

What are the modules?

- Word alignment (Giza++)
- Translation Modelling (Moses)
- Language Modelling (SRILM)
- Decoding (Moses)

System Extensions

Word Alignment

- Word Packing

Translation Modelling

- EBMT/Marker-based chunking
- Parallel treebank phrase extraction
- Supertagging

Language Modelling

- Supertagging

Resources

There are three main resources that will assist you when using the system:

- **MT Wiki** - http://mt-server-1.computing.dcu.ie/dokuwiki/doku.php
- **Makefile** - maia:/usr/local/share/matrex/Makefile
- Ask someone

Maintaining and Upgrading the System

Things that need to be done

- Upgrade core components – Moses, Giza++, SRILM, evaluation scripts
- Integrate new research into the system and document
  - source-context features
  - word packing, supertagging
  - subtree aligner
  - case/punctuation restoration techniques
- Update, upgrade and fully document the Makefile (in the wiki)
- Instructions for local installation

Maintaining and Upgrading the System

Things that could to be done

- Reordering model for EBMT chunks
- Extended compatibility with advanced Moses features
- Parallelisation


"That's all Folks!"