# Automatic Induction of Transfer Rules using Aligned Bilingual Corpora

Yvette Graham

National Center for Language Technology
School of Computing
Dublin City University

May 31 2006

# BiText Project

Project Aims

- Automatically generate a number of aligned bilingual corpora (bi-texts)
- Acquire existing bilingual corpora.
- Automatically annotate the bi-texts with LFG c-structures and f-structures.
- Automatically induce transfer rules from the corpora.

Traditionally, transfer rules for machine translation have been hand-coded - takes time and resources.

Automatically inducing such transfer rules from large bilingual corpora could provide an efficient way of acquiring such transfer rules.

# Transfer Rules for MT

SL sentence parsed and f-structure achieved.

Transfer Rule database is searched for a matching rule

- The SL sentence f-structure must match the LHS of the transfer rule.
- If multiple rules match, then the most specific and most likely one is selected.

Recursively apply transfer rules to any parts of the f-structure not taken care of by the initial rule.

Perhaps adding a default rule, for cases in which there is no matching transfer rule for an given f-structure.

Adding this information to the RHS of the initial rule produces the TL f-structure.

This f-structure is then used to generate the TL sentence.

# Some advantages of this approach to MT

The f-structure notation allows us to infer desirable transfer rules for f-structure relations that correspond to non-adjacent words and thus would not be discovered in the string-based world.

For example:

- **Es gibt ...**
  There is ...
  (lit. *It gives ...*)
- **Es scheint ... zu geben.**
  There seems [...] to be.
  (lit. *It seems ... to give.*)

- - Riezler, Maxwell (2006)

Transfer rules allow for mapping between different linguistic types.

For example, extraction of the phrase pair:

- **zutiefst dankbar ⇒ a deep appreciation**

Filtering of f-structure phrases based on consistency of linguistic types would find this mapping invalid because of the incompatibility of types A and N for adjective *dankbar* and nominal *appreciation*.

- - Riezler, Maxwell (2006)

# Automatically Inducing Transfer Rules

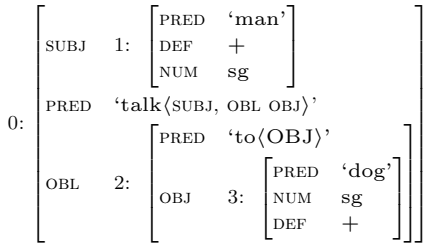
- F-structure annotated bilingual corpora are used to extract transfer rules for a given language pair.
- Relevant sentences and their translations are extracted from such corpora and analysed.
- The similarities between the SL f-structures are examined
  $\Rightarrow$ generalisation about the structure of *the sentence to be translated* and form the LHS of the transfer rule.
- The similarities amongst the TL f-structures are examined
  $\Rightarrow$ a generalisation about the structure of *the translation of the sentence* and forms the RHS of the rule.
- This results in a transfer rule meaning:
  given a SL sentence whose f-structure matches the LHS of the rule, the f-structure of its translation is likely to be that of the RHS of the rule.

# Example: Automatically Inducing Transfer Rule for the verb "*talk*" from English to French

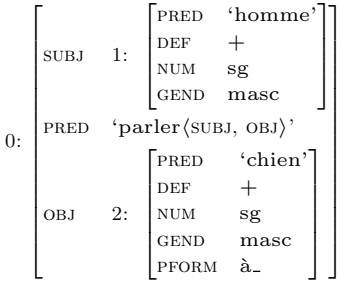
- Search the English side of all available English/French bi-text corpora for sentences in which the main predicate is *talk*.
- Retrieve both the English sentences and their matching French sentences.
- Examine the f-structures of the sentences looking for regularities amongst the English f-structures and then similarities amongst the f-structures of their translations.
- For Example: the search of the corpora could return the following three sentence pairs
    1. The man talks to the dog. *L'homme parle au chien.*
    2. He talks to Mary all day. *Il parle à Marie toute la journée.*
    3. The president talks to the nurse. *Le président parle à l'infirmière.*

# F-Structures for Sentence Pair 1

The man talks to the dog.




L'homme parle au chien.



# F-Structures as Equations: Sentence Pair 1

## English F-Structure:

$$\left\{\begin{array}{l}\textsc{subj}\ (0,\ 1)\\ \textsc{pred}\ (0,\ \text{`talk}\langle\textsc{subj},\ \textsc{obl}\ \textsc{obj}\rangle\text{'})\\ \textsc{obl}\ (0\ ,2)\\ \textsc{pred}\ (1,\ \text{`man'})\\ \textsc{def}\ (1,\ +)\\ \textsc{num}\ (1,\ \text{sg})\\ \textsc{pred}\ (2,\ \text{`to}\langle\textsc{obj}\rangle\text{'})\\ \textsc{obj}\ (2,\ 3)\\ \textsc{pred}\ (3,\ \text{`dog'})\\ \textsc{num}\ (3,\ \textsc{sg})\\ \textsc{def}\ (3,\ +)\end{array}\right\}$$

## French F-Structure:

$$\left\{\begin{array}{l}\textsc{subj}\ (0,\ 1)\\ \textsc{pred}\ (0,\ \text{`parler}\langle\textsc{subj},\ \textsc{obj}\rangle\text{'})\\ \textsc{obj}\ (0\ ,2)\\ \textsc{pred}\ (1,\ \text{`homme'})\\ \textsc{def}\ (1,\ +)\\ \textsc{num}\ (1,\ \text{sg})\\ \textsc{gend}\ (1,\ \text{masc})\\ \textsc{pred}\ (2,\ \text{`chien'})\\ \textsc{def}\ (2,\ +)\\ \textsc{num}\ (2,\ \text{sg})\\ \textsc{gend}\ (2,\ \textsc{masc})\\ \textsc{pform}\ (2,\ \text{à\_})\end{array}\right\}$$

# Linking the Content of the Two F-Structures

- Examine the values of the predicates of the English f-structure
- Translate each of the values using a bilingual dictionary or word list
- Search for the translated predicates in the French F-structure
- Record links between the f-structures

1. Predicate values of the english f-structure

$$\left\{\begin{array}{l}\text{PRED } (0, \text{`talk}\langle\text{SUBJ, OBL OBJ}\rangle\text{'}) \\ \text{PRED } (1, \text{`man'}) \\ \text{PRED } (2, \text{`to}\langle\text{OBJ}\rangle\text{'}) \\ \text{PRED } (3, \text{`dog'})\end{array}\right\}$$

2. Dictionary Look-Up

talk $\Rightarrow$ parler

man $\Rightarrow$ homme

to $\Rightarrow$ à

dog $\Rightarrow$ chien

3. Translations in the french f-structure predicate values:

$$\left\{\begin{array}{l}\text{PRED } (0, \text{`parler}\langle\text{SUBJ, OBJ}\rangle\text{'}) \\ \text{PRED } (1, \text{`homme'}) \\ \text{PRED } (2, \text{`chien'})\end{array}\right\}$$

4. Resulting links between f-structures:

| | |
|---|---|
| 0 $\Rightarrow$ 0 | write as variable A |
| 1 $\Rightarrow$ 1 | write as variable B |
| 3 $\Rightarrow$ 2 | write as variable C |

# Linking the Content of the F-Structures

English F-Structure:

$$\left\{\begin{array}{l}\textsc{subj}\ (\mathrm{A},\ \mathrm{B})\\ \textsc{pred}\ (\mathrm{A},\ \text{`talk}\langle\textsc{subj},\ \textsc{obl}\ \textsc{obj}\rangle\text{'})\\ \textsc{obl}\ (\mathrm{A}\ ,2)\\ \textsc{pred}\ (\mathrm{B},\ \text{`man'})\\ \textsc{def}\ (\mathrm{B},\ +)\\ \textsc{num}\ (\mathrm{B},\ \mathrm{sg})\\ \textsc{pred}\ (2,\ \text{`to}\langle\textsc{obj}\rangle\text{'})\\ \textsc{obj}\ (2,\ \mathrm{C})\\ \textsc{pred}\ (\mathrm{C},\ \text{`dog'})\\ \textsc{num}\ (\mathrm{C},\ \textsc{sg})\\ \textsc{def}\ (\mathrm{C},\ +)\end{array}\right\}$$

$\Rightarrow$

French F-Structure:

$$\left\{\begin{array}{l}\textsc{subj}\ (\mathrm{A},\ \mathrm{B})\\ \textsc{pred}\ (\mathrm{A},\ \text{`parler}\langle\textsc{subj},\ \textsc{obj}\rangle\text{'})\\ \textsc{obj}\ (\mathrm{A}\ ,\mathrm{C})\\ \textsc{pred}\ (\mathrm{B},\ \text{`homme'})\\ \textsc{def}\ (\mathrm{B},\ +)\\ \textsc{num}\ (\mathrm{B},\ \mathrm{sg})\\ \textsc{gend}\ (\mathrm{B},\ \mathrm{masc})\\ \textsc{pred}\ (\mathrm{C},\ \text{`chien'})\\ \textsc{def}\ (\mathrm{C},\ +)\\ \textsc{num}\ (\mathrm{C},\ \mathrm{sg})\\ \textsc{gend}\ (\mathrm{C},\ \textsc{masc})\\ \textsc{pform}\ (\mathrm{C},\ \text{à\_})\end{array}\right\}$$

# Examine the left side of the rules

LHS for Sentence Pair 1:

**SUBJ (A, B)**
**PRED (A,** 'talk⟨SUBJ, OBL OBJ⟩')
**OBL (A ,2)**
PRED (B, 'man')
DEF (B, +)
NUM (B, sg)
**PRED (2,** 'to⟨OBJ⟩')
**OBJ (2, C)**
PRED (C, 'dog')
NUM (C, SG)
DEF (C, +)

LHS for Sentence Pair 2:

**SUBJ (A, B)**
**PRED (A,** 'talk⟨SUBJ, OBL OBJ⟩')
**OBL (A ,2)**
ADJUNCT (A, D)
PRED (B, 'pro')
CASE (B, nom)
GEND (B, male)
NUM (B, sg)
**PRED (2,** 'to⟨OBJ⟩')
**OBJ (2, C)**
PRED (C, 'Mary')
PRED (D, 'day')
SPEC (D, 5)
QUANT (5, E)
PRED (E, 'all')

LHS for Sentence Pair 3:

**SUBJ (A, B)**
**PRED (A,** 'talk⟨SUBJ, OBL OBJ⟩')
**OBL (A ,2)**
PRED (B, 'president')
DEF (B, +)
NUM (B, sg)
**PRED (2,** 'to⟨OBJ⟩')
**OBJ (2, C)**
PRED (C, 'nurse')
NUM (C, SG)
DEF (C, +)

The common equations among the set of f-structures gives us the LHS of the transfer rule:

**SUBJ (A, B)**
**PRED (A,** 'talk⟨SUBJ, OBL OBJ⟩')
**OBL (A ,2)**
**PRED (2,** 'to⟨OBJ⟩')
**OBJ (2, C)**

# Examine the right side of the rules

RHS for Sentence Pair 1:

**subj (A, B)**
**pred (A, 'parler⟨subj, obj⟩')**
**obj (A ,C)**
PRED (B, 'homme')
DEF (B, +)
NUM (B, sg)
GEND (B, masc)
PRED (C, 'chien')
DEF (C, +)
NUM (C, sg)
GEND (C, MASC)
PFORM (C, à_)

RHS for Sentence Pair 2:

**subj (A, B)**
**pred (A, 'parler⟨subj, obj⟩')**
**obj (A, C)**
ADJUNCT (A ,D)
PRED (B, 'pro')
CASE (B, nom)
NUM (B, sg)
GEND (B, masc)
PRED (C, 'Marie')
PRED (D, 'journee')
QUANT (D, E)
SPEC (D, 5)
PRED (E, 'tout')
DET (5, 6)
PRED (6, 'le')

RHS for Sentence Pair 3:

**subj (A, B)**
**pred (A, 'parler⟨subj, obj⟩')**
**obj (A ,C)**
PRED (B, 'president')
DEF (B, +)
NUM (B, sg)
GEND (B, masc)
PRED (C, 'infirmiere')
DEF (C, +)
NUM (C, sg)
GEND (C, FEM)
PFORM (C, à_)

The common equations among the set of f-structures gives us the RHS of the transfer rule:

**subj (A, B)**
**pred (A, 'parler⟨subj, obj⟩')**
**obj (A, C)**

# Resulting Transfer Rule

$$\left\{\begin{array}{l}\text{SUBJ (A, B)}\\ \text{PRED (A, `talk}\langle\text{SUBJ, OBL OBJ}\rangle\text{')}\\ \text{OBL (A ,2)}\\ \text{PRED (2, `to}\langle\text{OBJ}\rangle\text{')}\\ \text{OBJ (2, C)}\end{array}\right\} \Rightarrow \left\{\begin{array}{l}\text{SUBJ (A, B)}\\ \text{PRED (A, `parler}\langle\text{SUBJ, OBJ}\rangle\text{')}\\ \text{OBJ (A, C)}\end{array}\right\}$$

Finally, replacing any remaining f-structure labels with variables:

$$\left\{\begin{array}{l}\text{SUBJ (A, B)}\\ \text{PRED (A, `talk}\langle\text{SUBJ, OBL OBJ}\rangle\text{')}\\ \text{OBL (A ,D)}\\ \text{PRED (D, `to}\langle\text{OBJ}\rangle\text{')}\\ \text{OBJ (D, C)}\end{array}\right\} \Rightarrow \left\{\begin{array}{l}\text{SUBJ (A, B)}\\ \text{PRED (A, `parler}\langle\text{SUBJ, OBJ}\rangle\text{')}\\ \text{OBJ (A, C)}\end{array}\right\}$$

# Current Project Status

Work to date:

- Compiled a list of available bilingual corpora.
- Developed a small trilingual corpus of sentences in German, English and French.
- Used XLE to parse the sentences producing their f-structures in prolog format.
- Manually compiled word-lists for the example sentences.

Currently working on:

- Implementing and testing different algorithms to extract transfer rules from the f-structures of aligned bilingual sentences for a given language pair.

# Some Future Work

Further develop the induction algorithm

Use the Human Centre Corpus to induce transfer rules and as validation for the induction algorithm

- French English (900 sent. approx)

Scale things up further to use a larger f-structure annotated corpus

Use the induced grammars developed by members of the LFG group to annotate existing bi-texts

Use these annotated bitexts to induce transfer rules