Multilingual and Parallel Corpora

# Alignment (ctd.)

Amir Zeldes

amir.zeldes@georgetown.edu

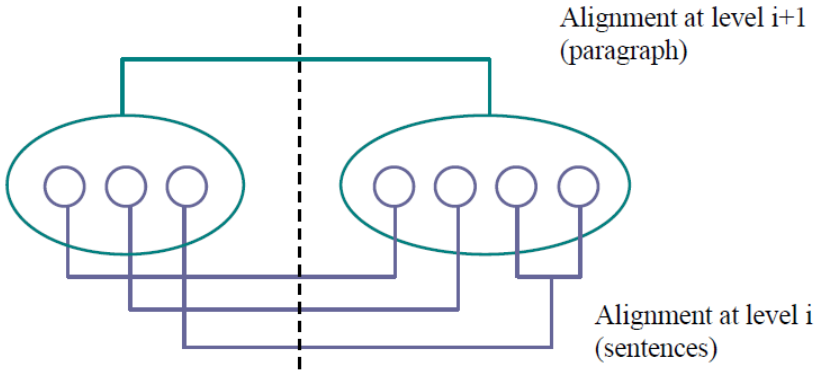
# Types of correspondence

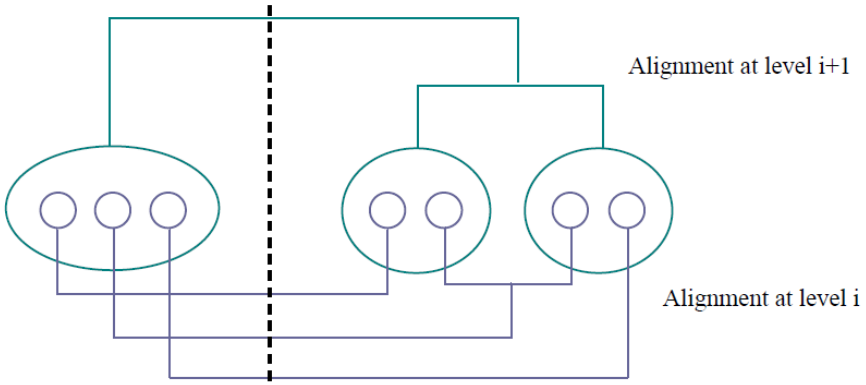
Non-linear correspondences:

- Null alignment
- Reordering
- Partial correspondence

# Types of multilevel alignment

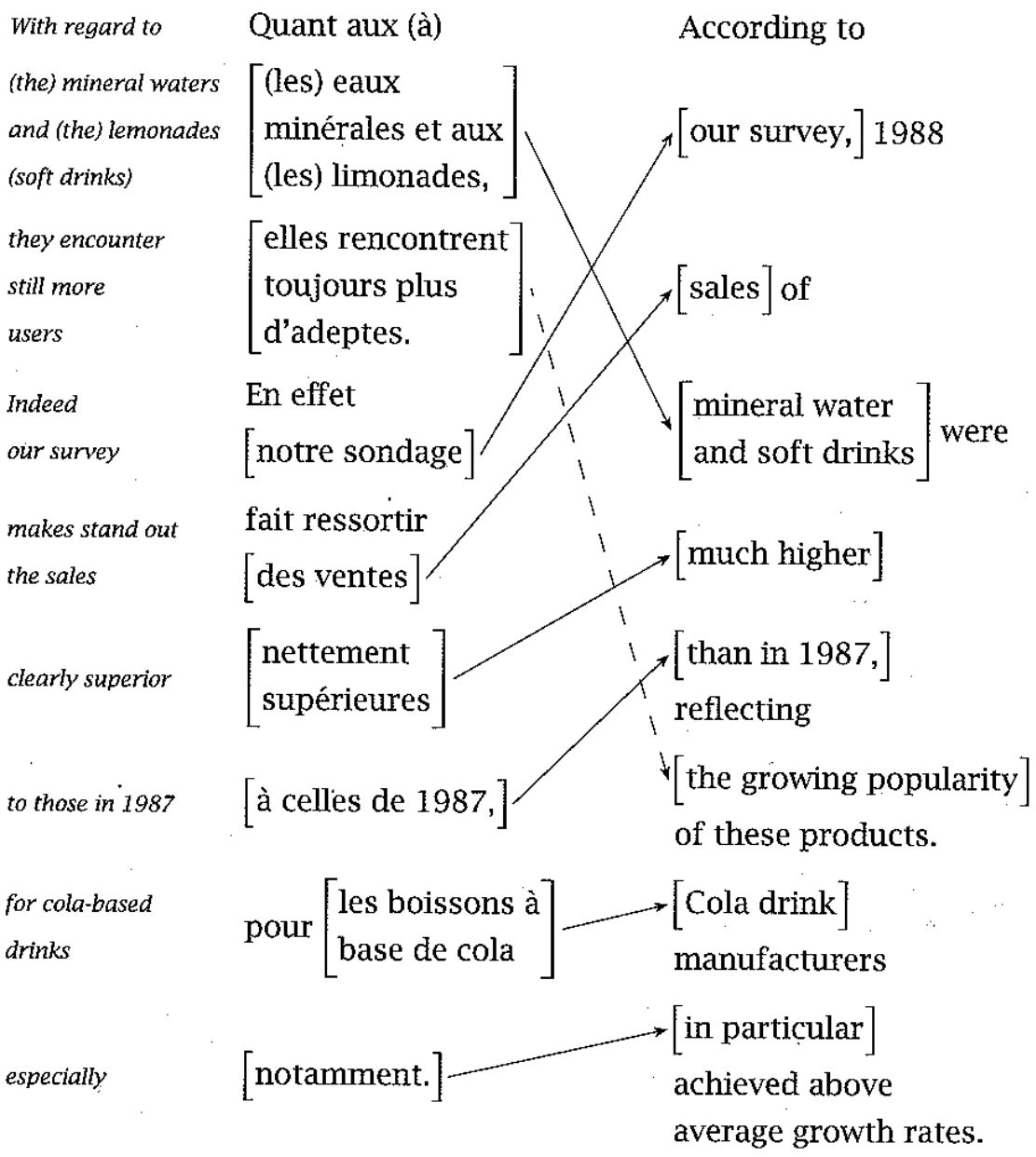
- We can divide multilevel alignments into 3 types:
  - Coherent
  - Compatible
  - Incompatible (hierarchy breaking, crossing edges)



# Getting automatic alignment

- What is realistically possible?
  - Sentence alignment
  - Length based approaches
- What do we want to have?


With regard to
Quant aux (à)
According to
(the) mineral waters and (the) lemonades (soft drinks)
(les) eaux minérales et aux (les) limonades,
our survey, 1988
they encounter still more users
elles rencontrent toujours plus d'adeptes.
sales of
Indeed
En effet
our survey
notre sondage
mineral water and soft drinks were
makes stand out
fait ressortir
the sales
des ventes
much higher
clearly superior
nettement supérieures
than in 1987, reflecting
to those in 1987
à celles de 1987,
the growing popularity of these products.
for cola-based drinks
pour les boissons à base de cola
Cola drink manufacturers
especially
notamment.
in particular achieved above average growth rates.

# Length – what do we count?

- Polish (7 words):
  *Mężczyzna w mundurze ocierał pot z czoła*
- German (12 words):
  *Der Mann in der Uniform wischte sich den Schweiß von der Stirn*
- English (10 words):
  *The man in uniform wiped the sweat from his forehead*

[Weiser Dawidek / P. Huelle]

# Length – what do we count?

- Letters? “Words”? Tokens?
- What do we do here:
  老王家有四个人
  *Old man Wang’s family consists of four people*

- In practice, it does not matter that much what we count…
- Main point: be consistent and compute the average **ratio** (e.g. De:En = 1.1 letters longer)

# Length based approaches

- Basic assumptions (Gale & Church 1993):
  - Sentences correspond to sentences of similar length
- Alignment candidates are often referred to as **beads**
- Probability of **1:1** bead alignment is usually highest
  - We can get a probability distribution of alignment types from some manually annotated data

# Bead alignment types

- Is there such a thing as p(1:2)?
  - In general?
  - For Eng:Spa?
  - For Eng:Spa novel?
  - For Cervantes?

# Bead alignment types

- What about crossing edges?
  - Option 1: have explicit crossing alignments
  - Option 2: have other types of n:n alignments:
    - 2:2
    - 3:3
    - But also: 2:3, 3:2 …
- What are the pros and cons?

# Length based approaches

- Basic assumptions (Gale & Church 1993):
  - Sentences correspond to sentences of similar length
- Alignment candidates are often referred to as **beads**
- Probability of **1:1** bead alignment is usually highest
  - We can get a probability distribution of alignment types from some manually annotated data
- Most often, **crossing edges** are rule out/ignored

# Length based approaches

- In practice, length based alignment is usually a two step process:
  - Paragraph alignment based on length (applicable to what genres?)
  - Sentence lengths within aligned paragraphs
- Works well for literal or 'near' translations
- 96%-99% correct for most benchmarks using **length in characters** (evaluation in Brown et al. 1991, comparison with approach using length in words)
- Danger: alignment can go haywire

# Lexical approaches

- Examine **content**
- Find **anchors** (certain alignment points)
- Stepwise alignment between anchors, until further anchors are found
- Search for word forms which consistently appear in parallel
- Some approaches restricted to using ‘content words’ (**POS tags**) or even proper names (tags and **capitalization**)

# Lexical approaches

Sicherer Anker. TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
Angela Merkel TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT Angela Merkel
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT Sicherer Anker.

Certain anchor. TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
Angela Merkel TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT Angela Merkel
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT Certain anchor.

# Non-identical anchors

- What can we do if the anchors are not identical?
  - Angela Merkel : Ангела Меркель
- Back to the Rosetta Stone:

# Quick exercise – anchor approach

- In the meantime , I should like to observe a minute ' s silence . You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka . One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam , who had visited the European Parliament just a few months ago .

- Σας καλώ να σηκωθείτε για αυτή την ενός λεπτού σιγή . Κυρία Πρόεδρε , επί ενός θέματος διαδικασίας . Θα έχετε ενημερωθεί από τον τύπο και την τηλεόραση ότι συνέβησαν ορισμένες εκρήξεις βομβών και φόνοι στη Σρι Λάνκα . Ένας από τους ανθρώπους που δολοφονήθηκαν πολύ πρόσφατα στη Σρι Λάνκα ήταν ο κ . Kumar Ponnambalam , ο οποίος είχε επισκεφθεί το Ευρωπαϊκό Κοινοβούλιο μόλις πριν λίγους μήνες

# Quick exercise – anchor approach

- In the meantime , I should like to observe a minute ' s silence .
- You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka .
- One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam , who had visited the European Parliament just a few months ago .

- Σας καλώ να σηκωθείτε για αυτή την ενός λεπτού σιγή .
- Κυρία Πρόεδρε , επί ενός θέματος διαδικασίας .
- Θα έχετε ενημερωθεί από τον τύπο και την τηλεόραση ότι συνέβησαν ορισμένες εκρήξεις βομβών και φόνοι στη Σρι Λάνκα .
- Ένας από τους ανθρώπους που δολοφονήθηκαν πολύ πρόσφατα στη Σρι Λάνκα ήταν ο κ .
- Kumar Ponnambalam , ο οποίος είχε επισκεφθεί το Ευρωπαϊκό Κοινοβούλιο μόλις πριν λίγους μήνες

# Quick exercise – anchor approach

- In the meantime , I should like to observe a minute ' s silence .
- You will be aware from the press and television that there have been a number of bomb explosions and killings in Sri Lanka .
- One of the people assassinated very recently in Sri Lanka was Mr Kumar Ponnambalam , who had visited the European Parliament just a few months ago .

- Σας καλώ να σηκωθείτε για αυτή την ενός λεπτού σιγή .
- Κυρία Πρόεδρε , επί ενός θέματος διαδικασίας .
- Θα έχετε ενημερωθεί από τον τύπο και την τηλεόραση ότι συνέβησαν ορισμένες εκρήξεις βομβών και φόνοι στη Σρι Λάνκα .
- Ένας από τους ανθρώπους που δολοφονήθηκαν πολύ πρόσφατα στη Σρι Λάνκα ήταν ο κ. Kumar Ponnambalam , ο οποίος είχε επισκεφθεί το Ευρωπαϊκό Κοινοβούλιο μόλις πριν λίγους μήνες

# Lexical Approaches

Certain anchor. TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
Angela Merkel TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT Angela Merkel
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT certain anchor.

**一定的 锚点.** TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT **安格拉·默克尔** TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT **安格拉·默克**
**尔** TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT **一定的 锚点.**

# Recognizing anchors

- Sometimes we have a list of anchor candidates
  - Lists of names
  - Transliteration rules
- But sometimes variation appears in the data
  - Can be language dependent, e.g. vowels in Semitic:
    - טרמפ
    - טראמפ
    - ...
- We need methods to recognize word **similarity**

# Edit Distance

- Recognize similarity between strings of characters
- Basic idea: try to turn one string into another in a minimal number of operations:
  - Delete a character: Cat**s** - Cat
  - Add a character: Schmid – Schmi**e**d
  - Replace a character: Schmi**d**– Schmi**t**

# Edit Distance

- What is the minimal number of operations?
  - Eng. Committee – Deu. Komitee
  - ➢ 3 (but: cost settings)
- Using Edit Distance and frequency analysis, we can get many more anchors to work with
  - Anchors will be less certain
  - But many anchors mitigate chance of alignment going haywire
  - Errors only occur between certain anchors

# Alignment and low resource languages

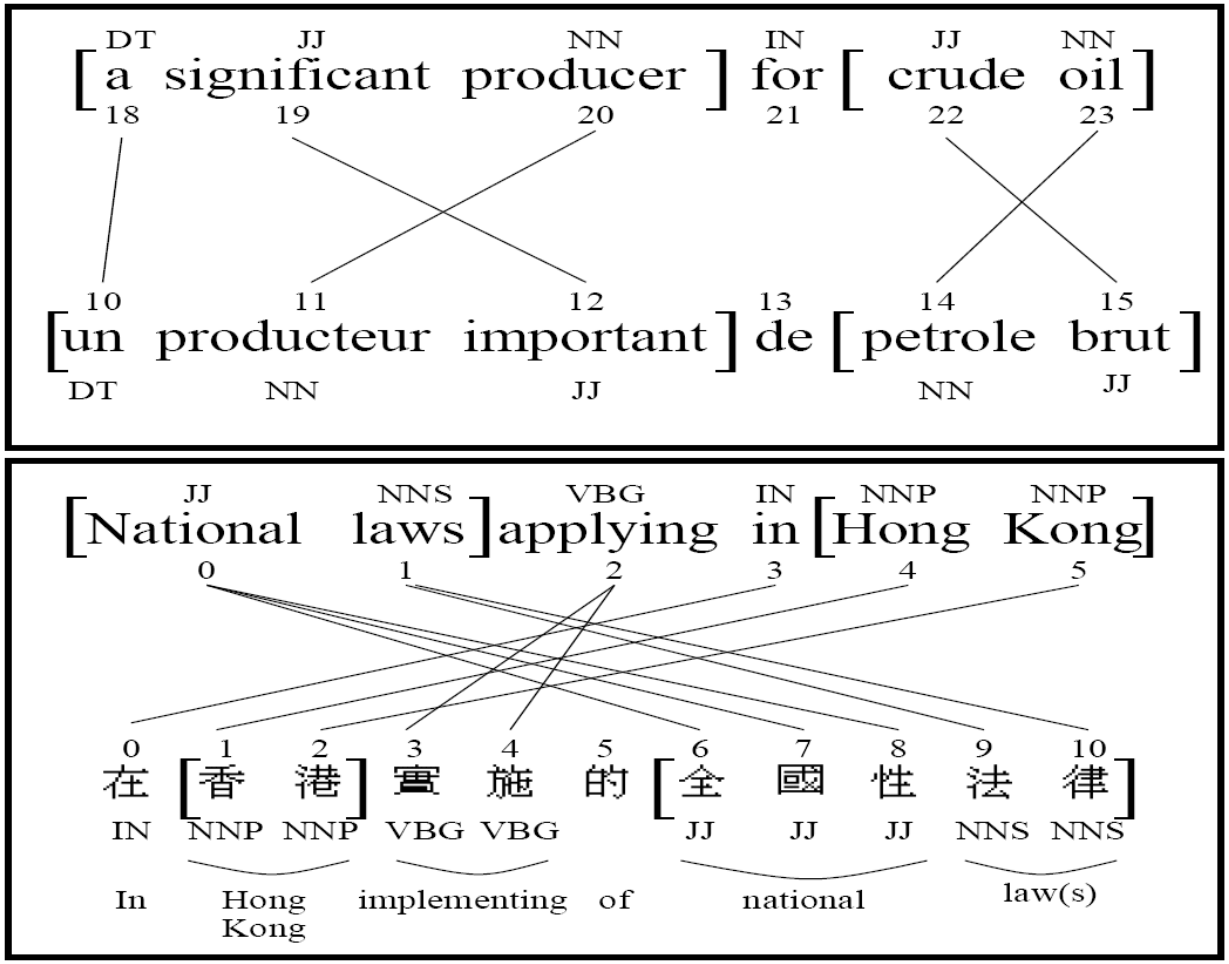
- For many languages we do not have automatic tools (part of speech taggers, syntactic parsers)
- But if we have a parallel corpus translating a high resource language (often English)…
- We can harness information from the parallel language!

# Annotation Projection

Approach pioneered by Yarowsky et al. 2001:

1. Add tags to high resource language (often English)
2. Get rough length based alignment
3. Assume aligned segments have same properties (=tags) as source segments
4. Train new tools on projected data

# Annotation Projection



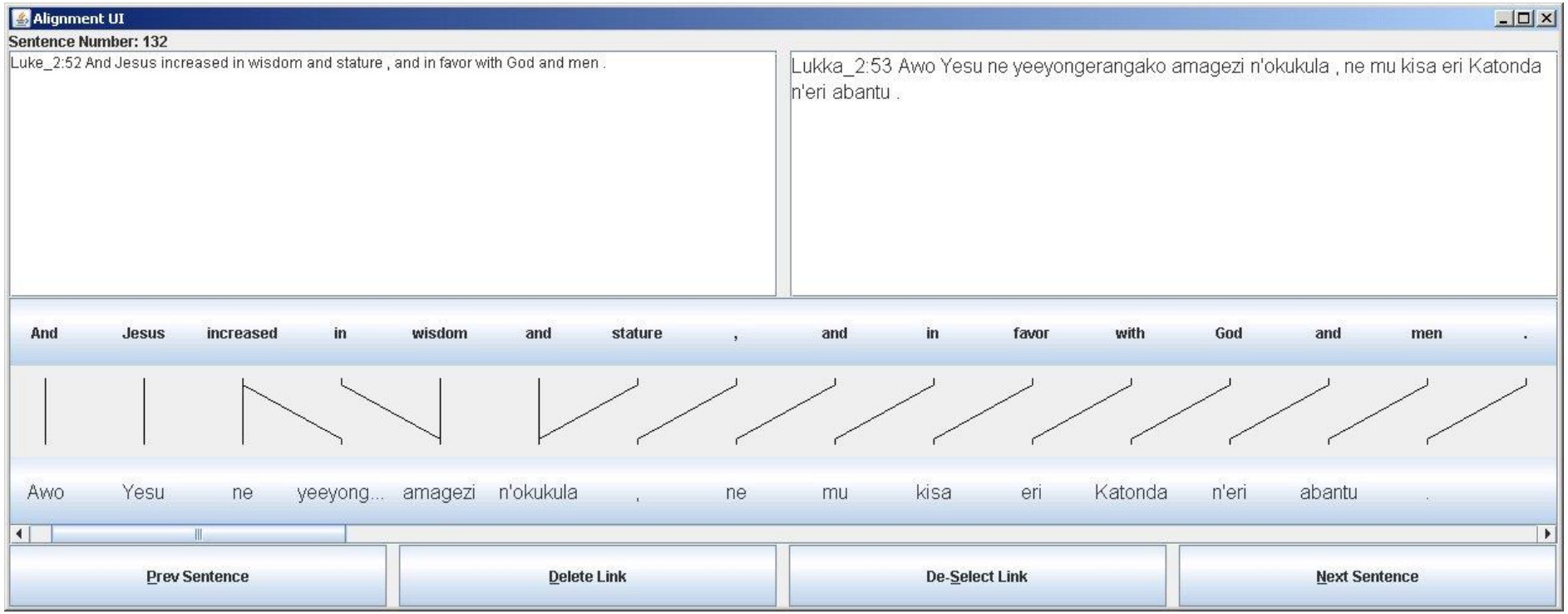

[Yarowsky/Ngai 2001]

# Annotation Projection

- Word alignment not always necessary:
  - Example – sentence type annotation
    - Declarative
    - Imperative
    - Polar questions
    - WH question
    - Hypothetical
    - …
  - Do these match up in your experience?

# For really useful tools we'd need…

- Word alignment!
- Can be trained manually - example: English - Luganda Parallel Corpus, http://aflat.org)



# For really useful tools we'd need…

- But much more often we rely on totally automatic, heuristic word alignment
  - How is this done?
  - Need to look at some machine translation algorithms
  - Overview - up next!

# Discussion

- Alignment is a non-trivial task
  - Is linear alignment sufficient?
  - What are the consequences of using automatic tools?
    - Length based approaches
    - Lexical/anchor based approaches
- What are the advantages and disadvantages of annotation projection?
  - What kinds of errors can we expect?
  - Are we constraining the TL annotation scheme?
  - Should we be worried about this and when?