Multilingual and Parallel Corpora

Alignment (ctd.)

Amir Zeldes

amir.zeldes@georgetown.edu

Types of correspondence

Simple linear correspondences:

- 1 to 1
- 1 to 2 (or more) – **split**
- 2 (or more) to 1 – **merge**

Non-linear correspondences:

- Null alignment
- Reordering
- Partial correspondence

Null alignment or null equivalence?

- Some points in the ST have no corresponding TT
- Easiest to identify when a clear translation is conceivable (something was just dropped)
  - When do we see **null equivalences**?
  - Are there also consistent **null correspondences**?
  - Are these **word** or **sentence** levels alignments?

Null equivalence in TL – word level

- *but he appears by his **own** confession to have been t he worse fordrink*
- *pero , según su **propia** confesión , estaba borracho .*

- *you have habitually underrated your **own** abilities*
- *siempre ha subestimado su habilidad **personal***

- *It was necessary to make a home **of my own** .*
- *Necesitaba un hogar .*

Null fertility

- Sometimes the TT **adds** text
- Another kind of null equivalence (in reverse)
- In Machine Translation sometimes referred to as ‘fertility’ (source material ‘spawns’ more target material)
- We’ll learn more about fertility when we look at MT in more detail

Examples – the word ‘Haus’ in German

Aus der Bibliothek - Drachenzucht für **Haus** und Hof - ist ein wenig veraltet , klar ,	Got this outta the library — Dragon Breeding for Pleasure and Profit — it ' s a bit outta date , o ' course
Jetzt waren nur noch drei Schüler übrig , deren **Haus** bestimmt werden mußte ."	And now there were only three people left to be sorted ."
wenn wir Drachen im Garten hinter dem **Haus** halten	if we ' re keeping dragons in the back garden

Null sentence alignment ('zero translation')

- Später, als er sprechen lernte, berichtete er, er sei Zeit seines Lebens in einem dunklen Kellerloch eingesperrt gewesen, er habe keinerlei Begriff von der Welt gehabt und nicht gewußt, daß es außer ihm noch andere Menschen gäbe… Er habe nicht gewußt, was ein Haus, ein Baum, was Sprache sei.

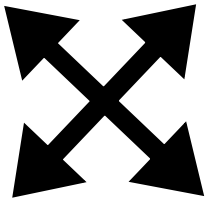
- Later he told of being locked in a dark cellar from birth. He had never seen another human being, a tree, a house before.

The Enigma of Kaspar Hauser
(see Munday 2001)

Later, when learned to speak, he reported he had been locked up his whole life in a dark cellar, he had not had any contact at all with the world and had not known that outside there were other people, because one slung food in to him, while he slept.
He did not know what a house, a tree, what language was.

Re-ordering

- Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes.
- En effet notre sondage fait ressortir des ventes nettement supérieures a celles de 1987.



- According to our survey, 1988 sales were much higher than in 1987.
- This reflects the growing popularity of mineral water and soft drinks.

[from Manning/Schütze 1999, 469]

Partial correspondence

- Greenwich

  Les envahisseurs venant du continent passaient par cette ancienne ville, par bateau ou par la Old Dover Road pour se rendre à la capitale

- Greenwich

  Vom Kontinent kommende Invasoren passierten sie auf ihrem Weg nach London entweder per Schiff oder über Straße Old Dover Road

[from Munday]

Partial correspondence

- Is it worthwhile to have a parallel corpus of parallel correspondences?
  - How can we build one?
  - What would we do with it?
  - What is the advantage over a real parallel corpus?
- Example:
  - http://opus.lingfil.uu.se/Wikipedia/de-en_sample.html

Discussion – Wikipedia Alignments

Translation

From Wikipedia, the free encyclopedia

This article is about language translation. For other uses, see Transla
"Translator" redirects here. For other uses, see Translator (disambiguation).
For article translations in Wikipedia, see Wikipedia:Translation.

Translation is the communication of the meaning of a source-language text by means of an equivalent target-language text.[1] While interpreting—the facilitating of oral or sign-language communication between users of different languages—antedates writing, translation began only after the appearance of written literature. There exist partial translations of the Sumerian *Epic of Gilgamesh* (ca. 2000 BCE) into Southwest Asian languages of the second millennium BCE.[2]

Translators always risk inappropriate spill-over of source-language idiom and usage into the target-language translation. On the other hand, spill-overs have imported useful source-language calques and loanwords that have enriched the target languages. Indeed, translators have helped substantially to shape the languages into which they have translated.[3]

Owing to the demands of business documentation consequent to the Industrial Revolution that began in the mid-18th century, some translation specialties have become formalized, with dedicated schools and professional associations.[4]

Because of the laboriousness of translation, since the 1940s engineers have sought to automate translation or to mechanically aid the human translator.[5] The rise of the Internet has fostered a world-wide market for translation services and has facilitated language localization.[6]

Translation studies systematically study the theory and practice of translation.[7]

תרגום

ערך זה עוסק בתרגום בין שפות. אם התכוונתם למשמעות אחרת, ראו תרגום (פירושונים).

תרגום הוא העברת מלל משפה אחת (שפת המקור) לשפה אחרת (שפת היעד), וזאת כדי שאנשים השולטים בשפת היעד, אך אינם שולטים בשפת המקור, יוכלו להבין מלל זה.

אדם העוסק בתרגום טקסט כתוב נקרא מתרגם, וזה העוסק בתרגום של דיבור, באופן סימולטני או מיד עם תום הדיבור, קרוי מתורגמן.

Übersetzung (Linguistik)

Unter **Übersetzung** versteht man in der Sprachwissenschaft einerseits die Übertragung eines (meist schriftlich) fixierten Textes von einer Ausgangssprache in eine Zielsprache, sie wird auch als „*Übersetzen*“ bezeichnet, anderseits versteht man darunter das Ergebnis dieses Vorgangs.

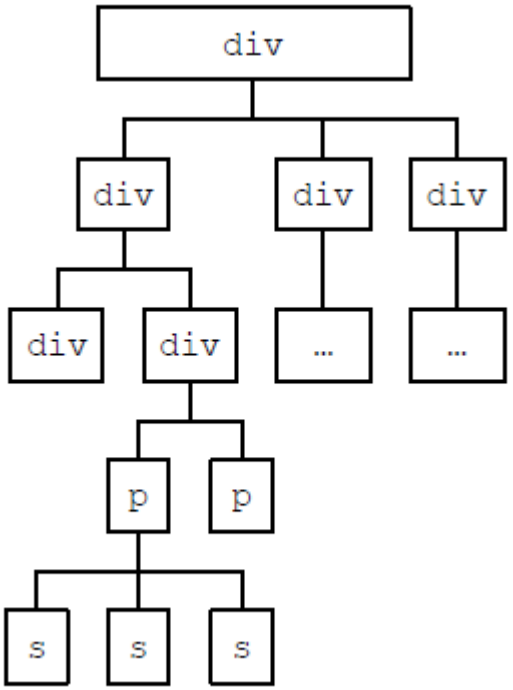
Zur besseren Unterscheidung wird das Produkt eines Übersetzungs- oder Dolmetschvorgangs (einer **Translation**) auch als *Translat* bezeichnet.

Die Übersetzung fällt gemeinsam mit dem Dolmetschen unter den Begriff Sprach- und Kulturmittlung (Translation). Der maßgebliche Unterschied zwischen Übersetzen und Dolmetschen liegt in der wiederholten Korrigierbarkeit des Translats. Wiederholte Korrigierbarkeit erfordert in aller Regel einen Zieltext, der in Schriftform oder auf einem Klangträger *fixiert* ist und somit wiederholt korrigiert werden kann, sowie einen in ähnlicher Weise fixierten Ausgangstext, den man wiederholt konsultieren kann. Liegt diese wiederholte Korrigierbarkeit vor, spricht man von einer Übersetzung. Ist jedoch der Ausgangstext oder der Zieltext *nicht fixiert*, weil er nur einmalig mündlich dargeboten wird, spricht man vom Dolmetschen. Veranschaulichen lässt sich das Prinzip anhand des Vom-Blatt-Dolmetschens: Hier liegt zwar der Ausgangstext schriftlich vor, aber der Zieltext ist nicht oder nur sehr eingeschränkt korrigierbar, da er nur gesprochen wird.

In der Sprachdidaktik wird häufig der Begriff *Mediation* verwendet. Im Unterschied zur *Translation* hebt der Begriff *Mediation* hervor, dass sich der Übersetzer oder Dolmetscher als *Mediator* in einer Vermittlungsposition zwischen zwei Personen befindet, die keine gemeinsame Sprache sprechen.

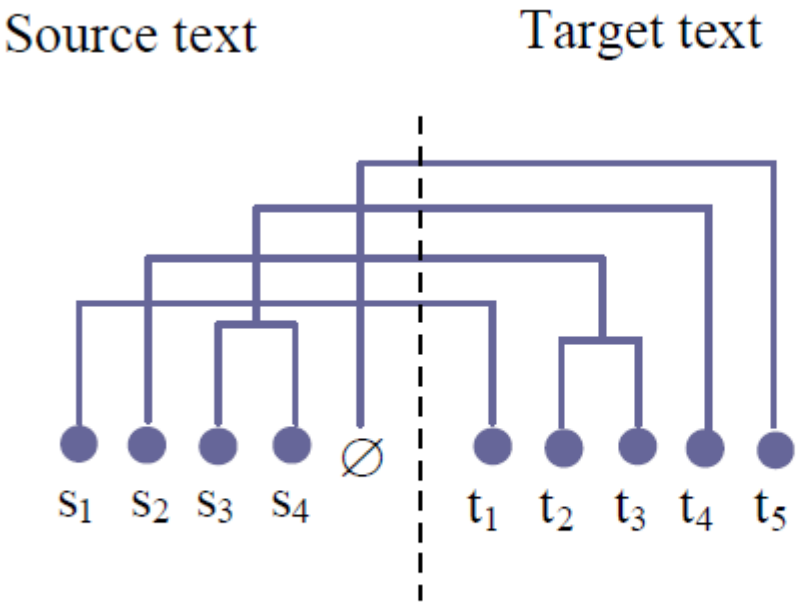
Multiple levels of alignment

- Romary & Bonhomme (2000) point out that texts generally follow hierarchical structure



Multiple levels of alignment

- In alignment, hierarchical levels are usually respected
- But there can easily be differences in quantity:



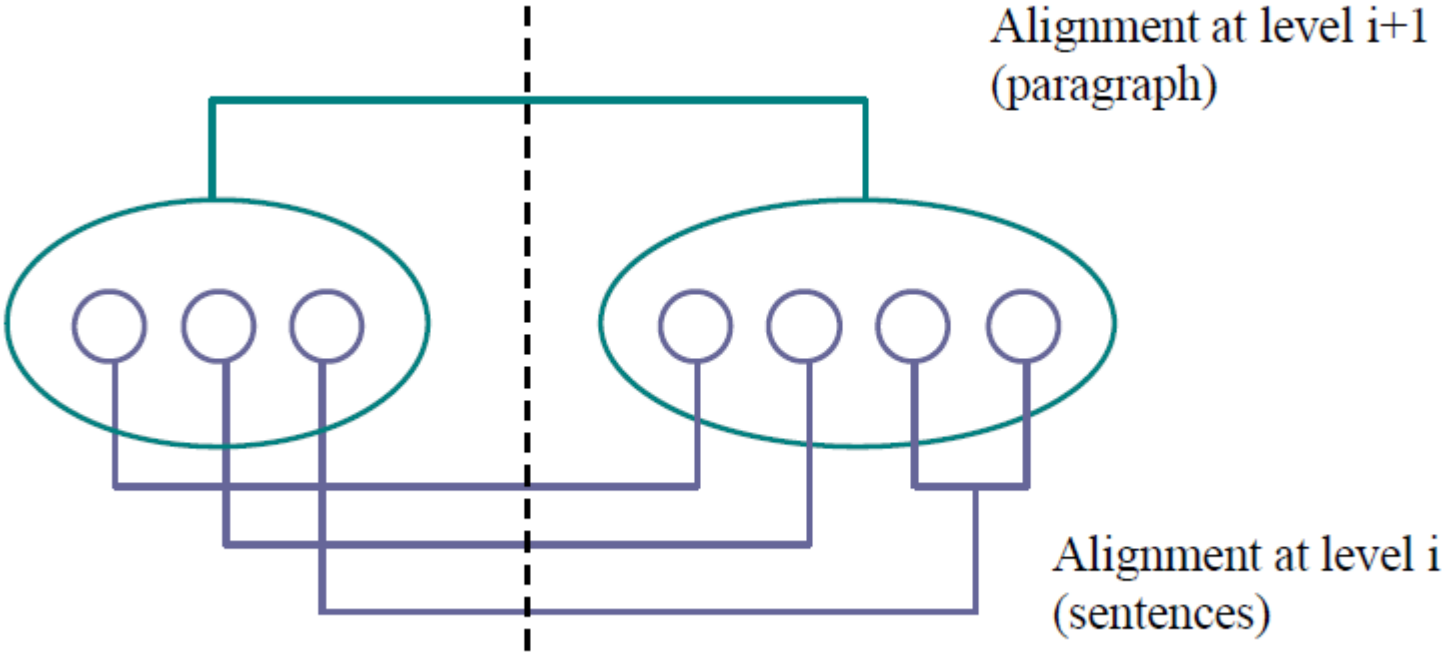
$$\text{Align}(S,T) = \{([s_1],[t_1]), ([s_2],[t_2,t_3]), ([s_3,s_4],[t_4]), (\varnothing,[t_5])\}$$

Types of multilevel alignment

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

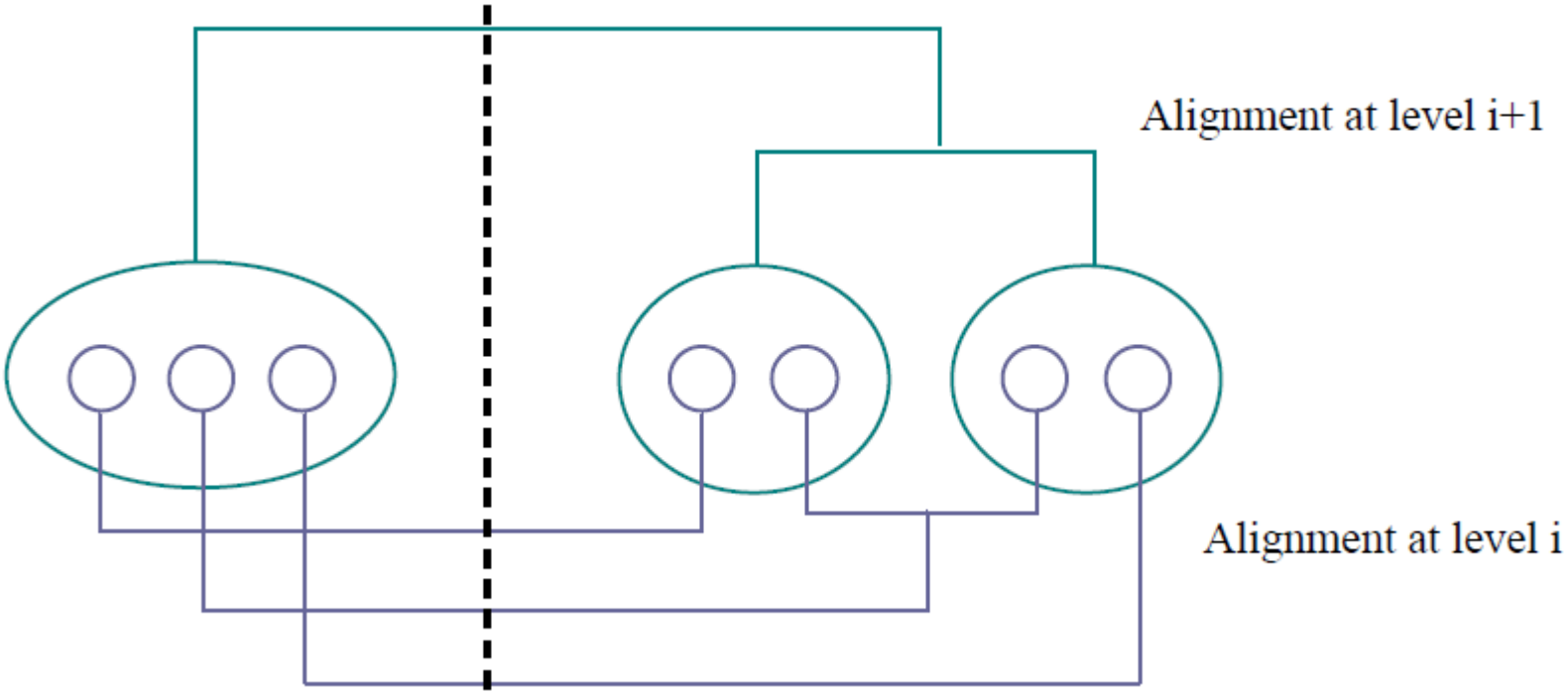
Types of multilevel alignment

- Coherent alignment



Types of multilevel alignment

- Compatible alignment (split across units)



How can we align parallel corpora?

- Corpora are often large (potentially very large, often anywhere from 1-1000 M tokens)
- Manual alignment is not possible, unless structured into the text (sections, etc.)
- Manual correction also impractical

➢ Fully automatic processes needed!

Alignment

- At a minimum, we want to align **sentences** (later: words)

DE	Das Thema ist so aktuell wie eh und je .
EN	The issue is as topical as it was before .
FR	Le thème est plus que jamais d' actualité .
	…

EuroParl, 11 languaged, 112 M English tokens

Length based approaches

- Basic assumptions (Gale & Church 1993):
  - Sentences correspond to sentences of similar length
- Is this true?

Languages and length

- 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]

Languages and length

- 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]

Next credit assignment

- Sentence alignment (due Wednesday)
  - Examine the two alternative versions of Greenwich text 1 in Canvas
  - Make a unit by unit alignment following the French-English example from Munday
  - For each alignment pair note as applicable:
    - Direct translation (nothing noteworthy)
    - Omission/fertility (elements with null alignment)
    - Shifts according to the shift approach
    - Occurrences of the 3 translation universals
  - Which alternative version is closer to V1?