Multilingual and Parallel Corpora

Translationese & Alignment

Amir Zeldes

amir.zeldes@georgetown.edu

Your translations!

- Literal of free?
  - Register
  - Intended audience
- Does logging influence your practice?
  - Did you actively avoid shifts?

Example – avoiding shifts

- Wilson (2009:178-179) discusses this example from Chesterman (2002):
  - *Le rapport biannuel sur la coordination des activités en faveur des PME et de l'artisanat rendra compte des progrès accomplis notamment grâce à l'établissement et à la comparaison de données sur le taux de participation des PME aux programmes communautaires tant en nombre de projets qu'en volume budgétaire et à l'introduction, le cas échéant, de mesures susceptibles d'augmenter la participation des PME.*

Example – avoiding shifts

- The twice-yearly report on the coordination of activities to assist SMEs and the craft sector will detail the progress achieved, particularly through compiling and comparing data on the participation rate of SMEs in Community programmes – in terms both of the number of projects and the budgetary volume involved – and through the introduction, where appropriate, of special measures to increase the participation of SMEs.

- Our success or failure will be measured by the twice-yearly report on action to help small businesses. This will show exactly how many of them are involved in Community programmes – both the number of projects, and the financial volume they represent. The report will also chart the impact of any special measures that might boost applications from small businesses.

Translation Universals

- Three processes (Baker 1996):
  - Explicitation
  - Simplification
  - Normalization (a.k.a. ‘levelling out’)

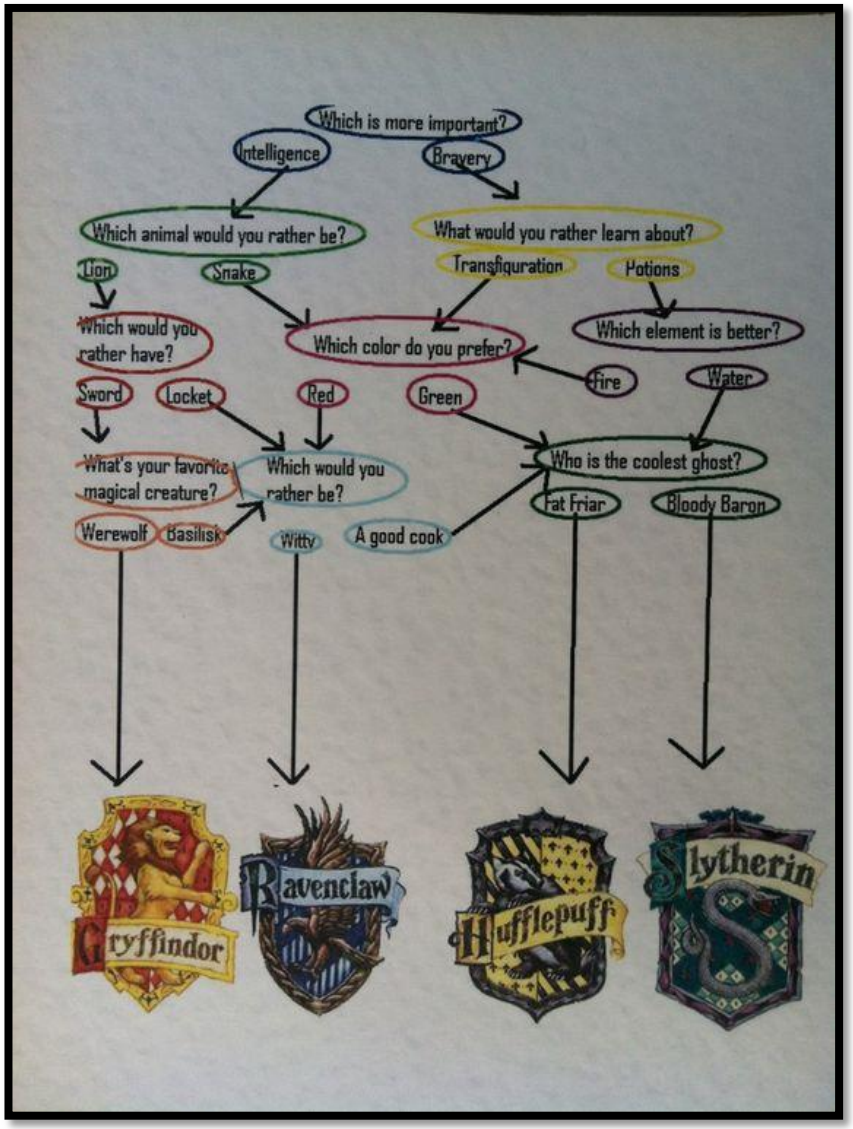
Explicitation

And now there were only three people left to be sorted.

Jetzt waren nur noch drei Schüler übrig , deren Haus bestimmt werden mußte .

Now were only still three pupils remaining, whose house determined become had-to

- sorted = into houses
- People = Schüler



Simplification

*but **there was no escaping** Dudley ' s gang , who visited the house every single day*

*doch Dudleys Bande , die das Haus Tag für Tag heimsuchte , **konnte er nicht entkommen***

Yet Dudley's gang, which the house day for day plagued could he not escape

- "Standard" sentences and simple constructions are preferred

Normalization / Levelling

- How can non-standard language be translated?
  - *Begbie was hard, but not so hard that he didn't shite it off twenty years in Saughton.*
  - *Y por duro que Begbie fuese , veinte años en prision no los iba a aguantar .*
  - *Begbie était dur , mais pas au point de se foutre de vingt ans de taule .*

Translation universal - consequences

- In practice, non-standard features are usually simply avoided
- Applies especially to dialectal forms
- Consequences:
  - TL texts are more similar to each other than SL originals
  - Certain meaning nuances are lost

Translationese (Gellerstam 1986)

- Translated texts have idiosyncratic properties compared to native ones
  - Some differences due to lexical frequencies
  - Related to cultural background (coffee vs. beer)
- But some differences are grammatical:
  - Answering with yes/no vs. verb
- Some involve idiomatic expressions: he said ***smiling***

Grammatical Translationese

- Typology from Santos (1996):
  - A > **B**, C, D – A structure has multiple equivalents, but one is preferred
  - A + obligatory > A + **optional** – an optional modification is frequent above and beyond chance compared to non-translate use in order to match ST
  - Vague(A/B) > **unmarked**(A') – if ambiguity cannot be preserved, the unmarked variant is taken
  - Complex(A&B) > A'+B' / A / B – complex meaning is separated into individual parts and translated as far as possible

Examples – complex meaning

- Santos (1996) – small parallel corpus (100K tokens)
- English : Portuguese, Fiction
  - Por. Imperfect > Eng. Progressive
    (*torcia as mãos > He was twisting hands*)
  - *He ran out of the house*
    > Ele saiu da casa a correr (no beginning)
    he left the house running
  - *> Ele correu para fora da casa* (no end)
    He ran to the outside of the house
  - *> Ele correu para fora da casa e saiu* (separated)
    He ran out of the house and left (successfully)

Identifying translations

Baroni & Bernardini (2005):

- Machine learning with a corpus of 2M tokens Italian + ~800K tokens of translations
- Over 86% automatic recognition of translated Italian
- Computer outperforms human judges

Discussion - Translationese

- Is translated language a variety in itself?
- Can/should we describe translated language independently of ‘normal’ language?
- What does this mean for training translators? (Think of the IoL criteria!)

Discussion - Translationese

- How does translationese affect corpus use for language comparison?
  - Multi-directional corpora – both languages as SL
  - Multiple translations of the same text?
- Testing hypotheses from parallel data in **comparable** corpora
- Comparison with non-corpus based approaches

Alignment

Plan

- What are the minimal and maximal units of translation?
- What can affect a translation and at what distance?
- How can we find translation units manually?
- How can we use computers to produce an alignment automatically?

Types of correspondence

- We have already seen numerous non 1:1 alignments
- What kind of correspondences are possible in practice?
- Which ones appear and how often?
- Does this depend on individual language pairs? Genres? Translators?

Types of correspondence

Simple linear correspondences:

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

1-1

- No two Ollivander wands are the same , just as no two unicorns , dragons , or phoenixes are quite the same .
- Keine zwei Ollivander - Stäbe sind gleich , ebenso wie kein Einhorn , Drache oder Phönix dem andern aufs Haar gleicht .

1-2

- No two Ollivander wands are the same , just as no two unicorns , dragons , or phoenixes are quite the same .

- Keine zwei Ollivander - Stäbe sind gleich.
- Es ist genau so, wie kein Einhorn , Drache oder Phönix dem andern aufs Haar gleicht .

2-1

- No two Ollivander wands are the same.
- It ‘s just like no two unicorns , dragons , or phoenixes are quite the same .

- Keine zwei Ollivander - Stäbe sind gleich , ebenso wie kein Einhorn , Drache oder Phönix dem andern aufs Haar gleicht .

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?

Next credit assignment

- Sentence alignment (due next Wednesday, 2/22)
  - 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?