Using Example-Based Machine Translation to translate DVD Subtitles

Marian Flanagan
Centre for Translation and Textual Studies
School of Applied Language and Intercultural Studies
Dublin City University, Ireland
marian.flanagan@dcu.ie

13 November 2009, Dublin City University

This research is funded by the Irish Council for the Humanities and Social Sciences (IRCHSS)

Presentation Outline

- Subtitling and Technology
- Aims of the Study
- Corpus-Based MT Approaches
- Experimental Design
- Results
- Conclusions and Suggestions for Future Work

Subtitling and Technology

- Audiovisual Translation (AVT) in Europe is a multi-million Euro industry
- Subtitling is part of this industry

- Why introduce technology into the domain of subtitling?
  - Digitalisation (arrival of DVD)
  - Increased pressures (time, costs)
  - Accessibility legislation

- Characteristics of subtitles

- Industry Collaboration: SMT example [Volk 2008]
  - Corpora
  - Swedish subtitling company
  - Text Shuttle (http://www.textshuttle.ch/main.php) [2009]

Review of the literature

- NHK (Japan) late 80s began generating automated subtitles for news broadcasts: RBMT → EBMT (2003)
- Popowich et al., (2000): RBMT
- O'Hagan (2003): TM and RBMT
- Piperidis et al., (2005): MUSA - Speech, TM and RBMT
- Melero et al., (2006): eTITLE - TM and RBMT
- Armstrong et al., (2006): EBMT
- Armstrong (2007): EBMT
- Volk and Harder (2007), Volk (2008), Hardmeier and Volk (2009): SMT

Aims of the Study

- Investigate quality of EBMT-generated subtitles
- End-user requires intelligible and acceptable subtitles

- We investigate if
  - increasing levels of source language repetitions between the test and training data
  - increasing the size of the corpus
  - decreasing the homogeneity of the corpus

  has a significant impact on the **intelligibility** and **acceptability** of EBMT-generated subtitles?

- Intelligibility is a necessary but not a sufficient condition for acceptability of subtitles

Data-Driven MT Approaches

- Corpora
  - Bilingually-aligned subtitling corpora
  - Large amounts of data
  - Genre specific (documentaries, DVDs etc)
  - Use of corpora such as Europarl (Koehn 2005)
- Industry Collaboration
  - Provide corpora
  - More easily aligned
  - Genre specific

EBMT in this Study

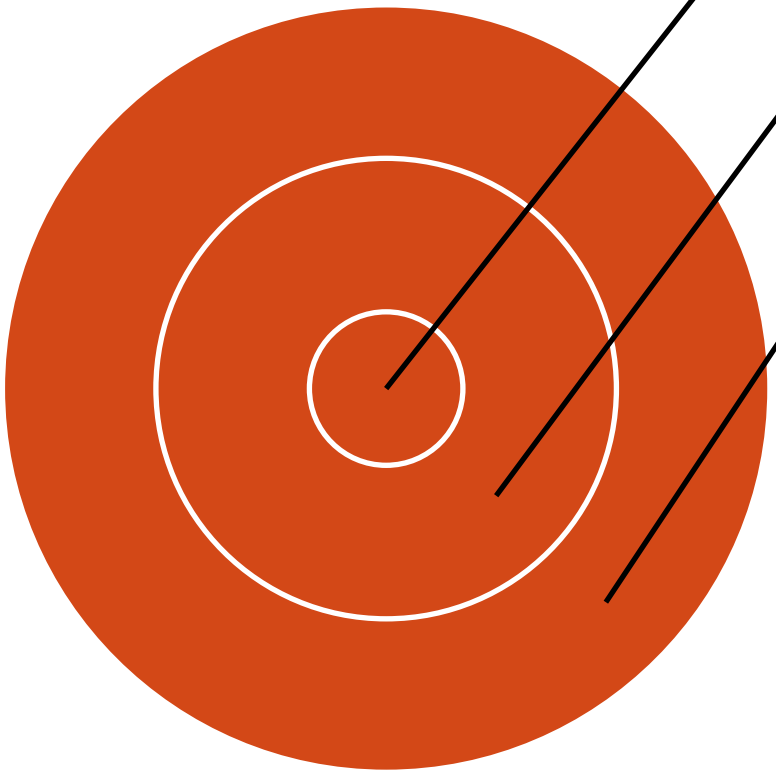
- MaTrEx system (2007)
  - Hybrid system
  - Marker Hypothesis
  - Phramer decoder

- Test Sets and Training Corpora
  - Six 2-minute clips from Harry Potter (23-36 subtitles per clip)
  - 3 training corpora (A, B, C)
  - Differ in size and in homogeneity, with Corpus A being the most homogeneous (fantasy)
  - Differ in the number of repeated source language (SL) segments they contain, with Corpus C, the largest corpus, containing the highest number of repeated SL segments

What's in Corpus A, Corpus B and Corpus C?

- Corpus A: 6997 aligned subtitles (taken from the first four Harry Potter movies)
- Corpus B: 11,342 aligned subtitles (Corpus A + subtitles taken from the Lord of the Rings trilogy)
- Corpus C: 42,331 aligned subtitles (Corpus B + subtitles taken from 25 DVDs from genres other than fantasy)

Training Corpora



Corpus A
(6,997)
Corpus B
(11,342)
Corpus C
(42,331)

Experimental Design

- Evaluation of automatically-generated subtitles
  - Multimodal texts vs. general texts
  - BLEU scores: de facto standard?
    - BLEU scores come second to human judgements of MT output, with automatic scores described as an imperfect substitute for human assessment of translation quality
    - Relationship between BLEU scores and subtitle quality?
  - Human evaluation
    - FEMTI (Framework for Evaluation of Machine Translation in ISLE) and Recipient Evaluation
    - Intelligibility (comprehensibility and readability) and acceptability (style and well-formedness)
    - End-users of the subtitles

Evaluation

- 44 German native speakers: **Corpus A (15), B (15), C (14)**
- All subjects had previously watched subtitles on DVD
- Half of the subjects had previous knowledge of Harry Potter
- Almost half of the subjects had formal training in linguistic issues
- Soundtrack of movie clips alternated between English (language known to evaluators) and Dutch (unknown language)
- Interview questionnaire: scale, open and closed questions
- Comprehensibility, style and observed errors ranked on scale, 1 (lowest) – 6 (highest)

Results/1: Quantitative

- Comprehensibility: Corpus B subtitles ranked highest (not statistically significant)

Corpus A	Corpus B	Corpus C
3.10	**3.35**	3.17

- Readability: Speed of Corpus C subtitles deemed the most suitable, and therefore more readable (statistically significant)

Is the speed of the subtitles suitable?		
Corpus A	Corpus B	Corpus C
81.1%	72.2%	**91.7%**

Results/2: Quantitative

- Style: Corpus B subtitles ranked highest in terms of style (statistically significant)

Corpus A	Corpus B	Corpus C
3.35	**3.87**	3.61

- Well-formedness: Seriousness of errors is not statistically significant, but the number of Class 1 errors noted in Corpus C subtitles is statistically significant, thus reducing the well-formedness

Corpus A	Corpus B	Corpus C
7	12	**30**

Results/1: Qualitative

- Comprehensibility: all corpora received negative comments, including

At times reading ‘normal words’ was problematic

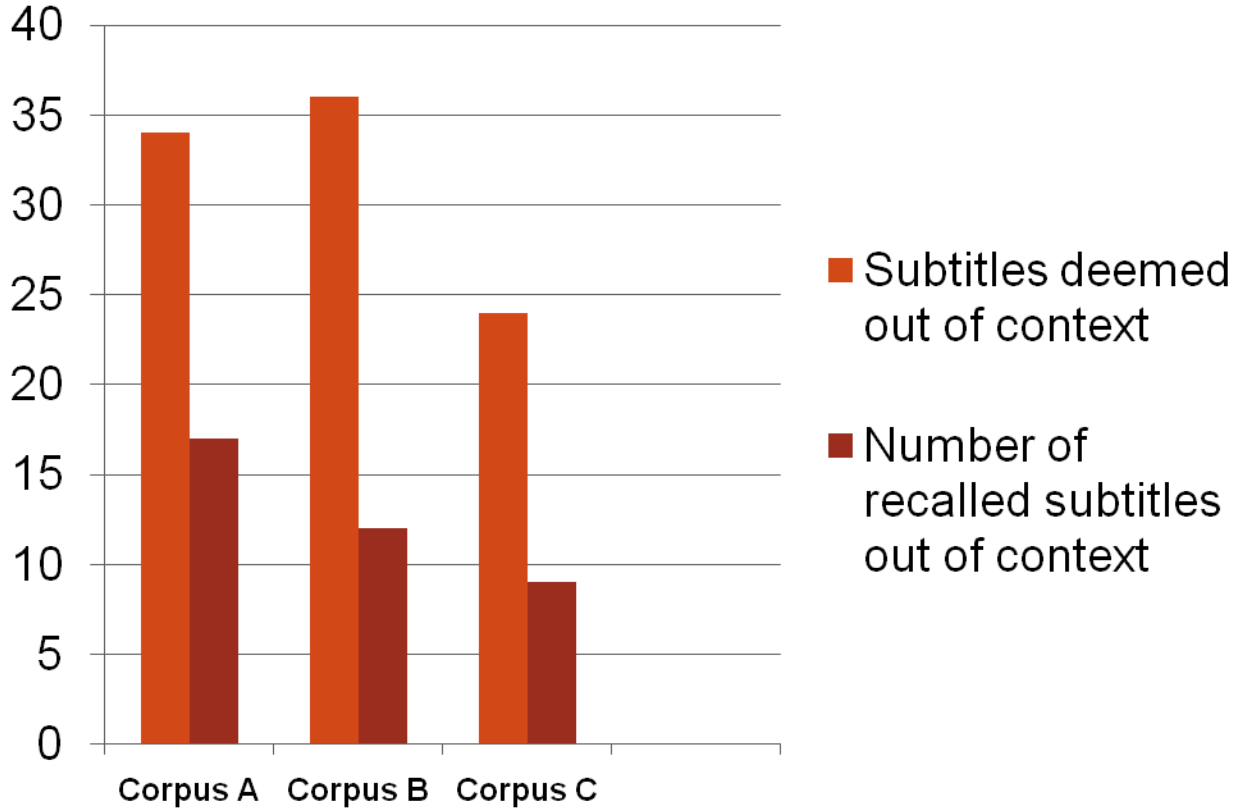
Some subtitles were ‘strange’, I couldn’t understand them

I could understand more from using the Dutch soundtrack rather than the subtitles

I only understood the meaning of the subtitle from using the English language soundtrack

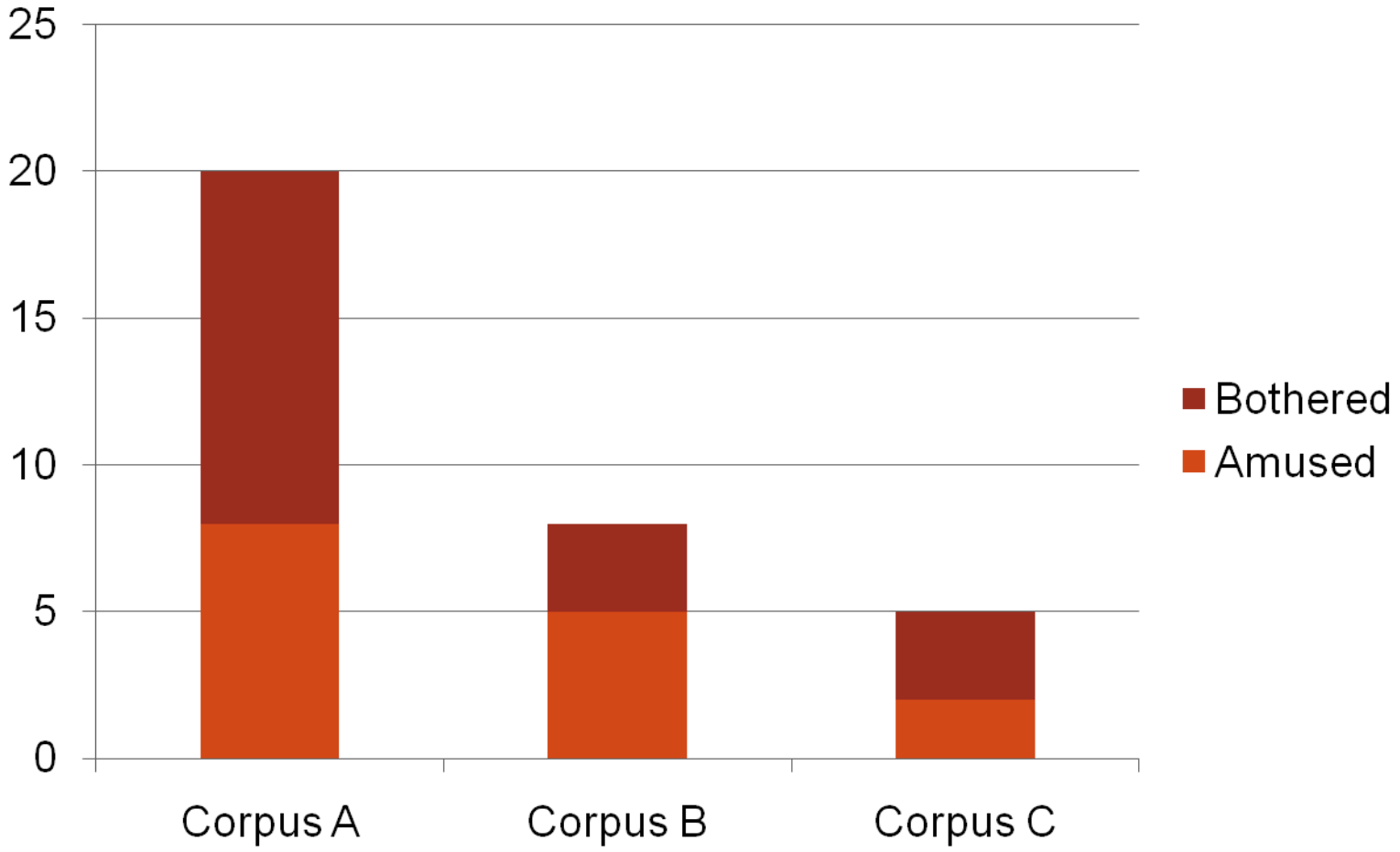
Results/2: Qualitative

- Readability: supports quantitative result



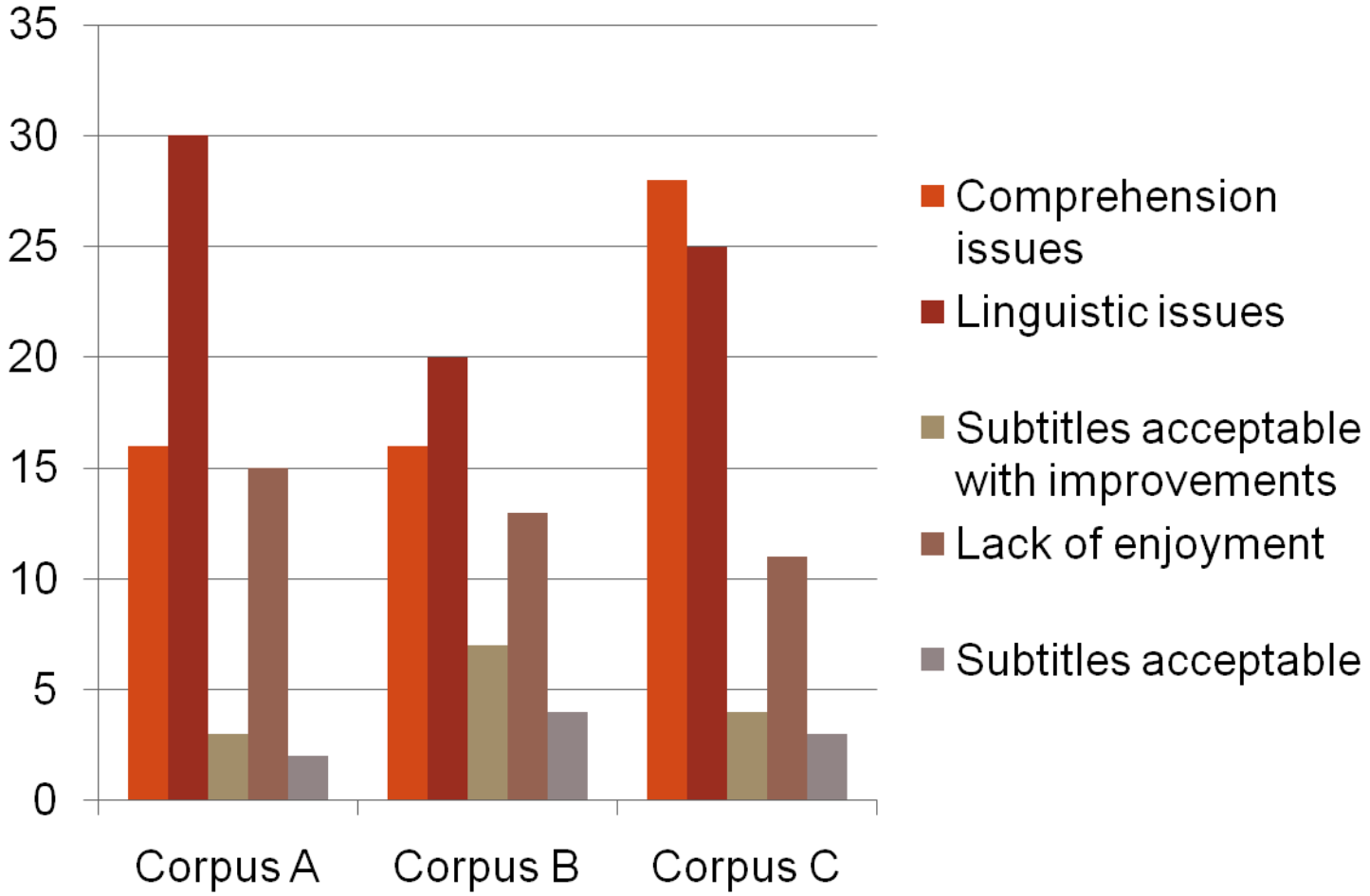
Results/3: Qualitative

- Style: supports quantitative result



Results/4: Qualitative

- Well-formedness



Results: Quantitative and Qualitative

- Combining qualitative and quantitative results:
  - the **readability** of machine-generated subtitles is improved if the size of the corpus is increased and with that the number of SL repetitions, and the corpus heterogeneity, in relation to the movie clips being subtitled (Corpus C)
  - On the other hand, the **comprehensibility** and **acceptability** of the subtitles is not improved given the same conditions (with Corpus B proving most successful given these metrics)

BLEU Scores

- English-German Harry Potter subtitles

	Corpus A	Corpus B	Corpus C
All 6 movie clips	25.97	**26.26**	26.11

- SMT (Volk 2008) Swedish – Danish

Crime Series	Comedy Series	Car Documentary
63.9	54.4	53.6

- SMT (Koehn 2005 and 2009)

	Swedish-Danish	English-German	Highest
Europarl	30.3	17.6	39.0 (pt-fr)
Acquis Communautaire	46.6	46.8	64.0 (fr-en)

Conclusions

- EBMT-generated subtitles are intelligible and acceptable in certain circumstances
- These findings are based on raw EBMT output. Automatically-generated subtitles would only ever serve as a draft for the subtitler
- BLEU scores and human judgements are somewhat related, however, this would need to be tested again – perhaps using an alternative metric
- Need for user reception studies in AVT
- Non-technical evaluation promotes interdisciplinary research

Suggestions for Future Work

- Pursue subtitle generation as a real-world application of EBMT – there is a real need for this service
- Empirical evaluation using
  - online modules
  - web service such as Amazon's Mechanical Turk
- Sharing of corpora: academia/industry
- Collaborative work with industry encouraging given the success to date of SMT
- Open source tools!

Questions?

Thank you for your attention