Machine Translation: broadening our horizons

Dorothy Kenny

SALIS / Centre for Translation and Textual Studies

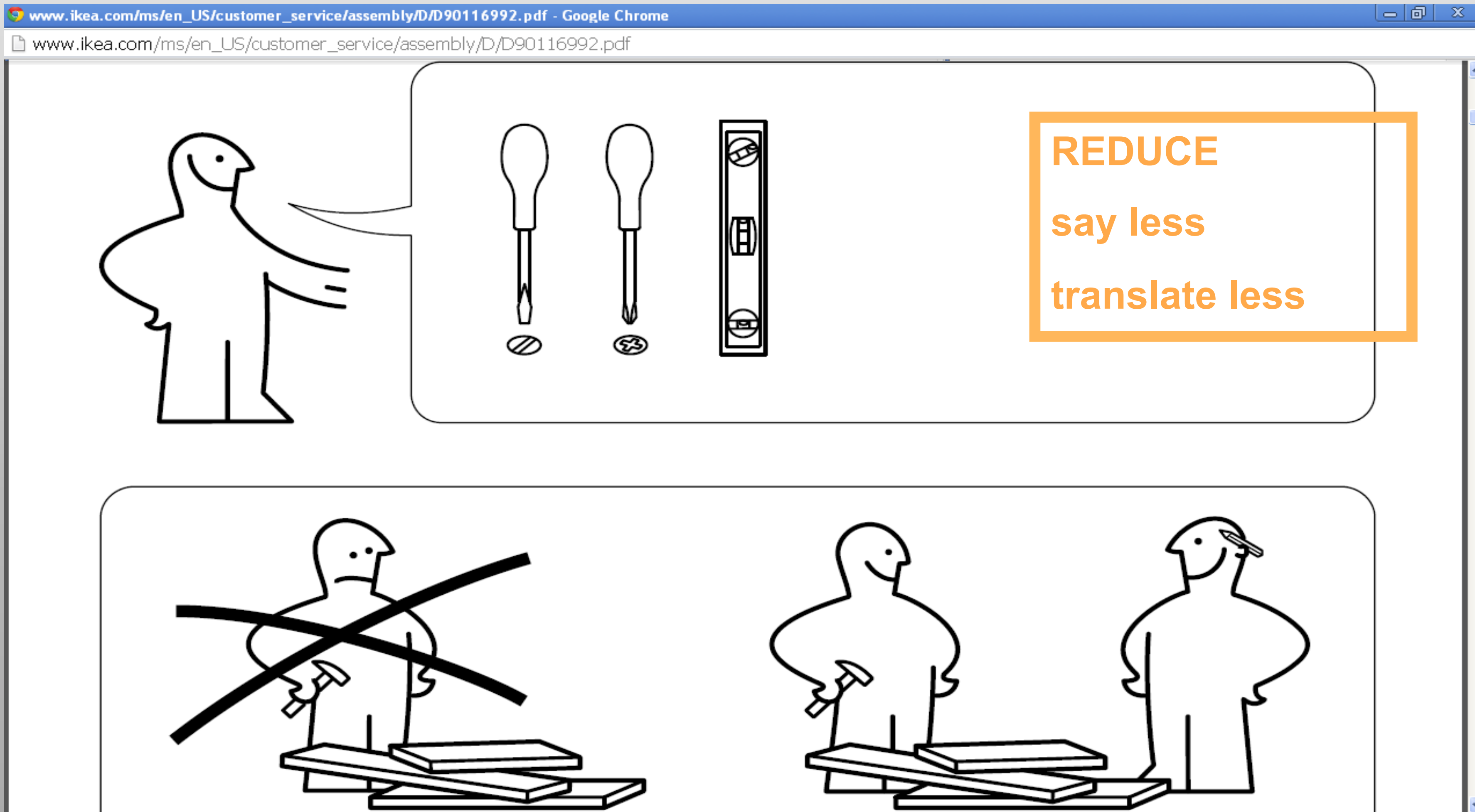
Dublin City University, Ireland

Durham 25 November 2015

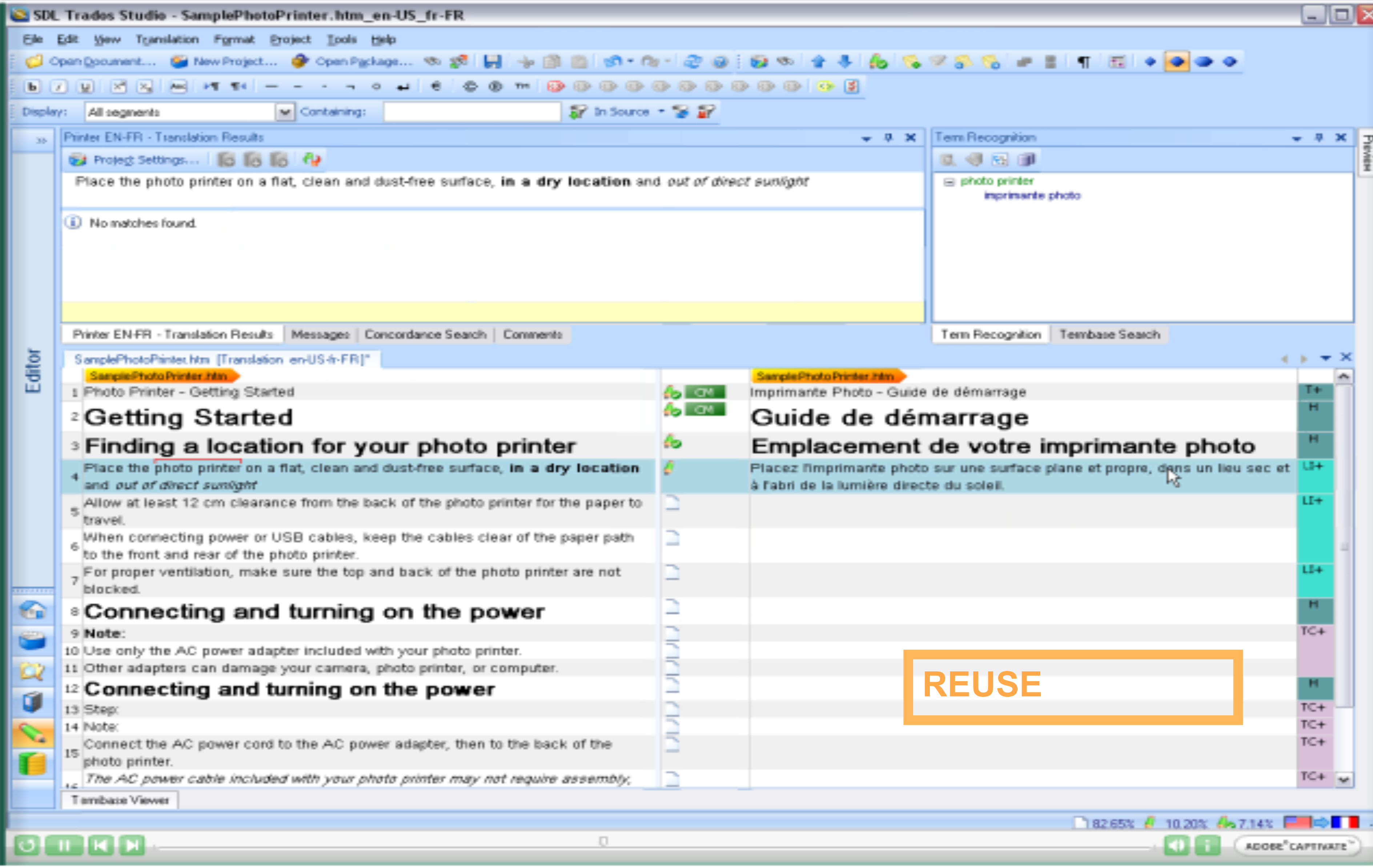
Outline

- Contemporary translation technology in three words
- Statistical Machine Translation
- Translation Studies reactions to SMT
- Counter-reactions: SMT at Dublin City University
- Ethics and the wider world

Translation Strategies



Translation Memory: Reuse



Statistical Machine Translation

Translations stored in translation memories can be reused.

They can also be recycled…

Example SMT Systems

Asia Online™

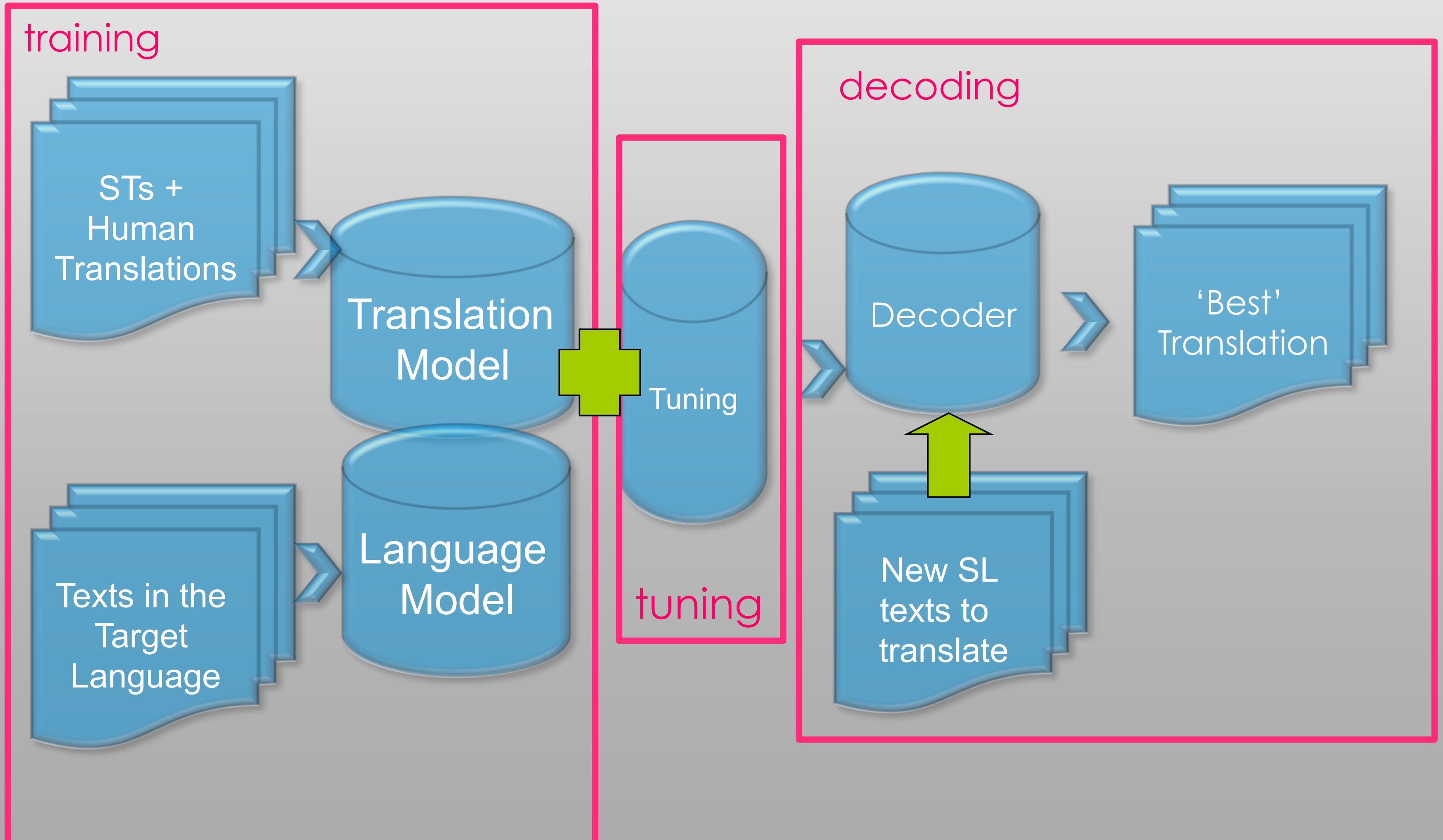
KantanMT.com

Microsoft Translator

Google translate

MT@ec

SMT Architecture



So what gets recycled?

- Phrase translations for den Vorschlag learned from the Europarl corpus:

English	$\phi(\bar{e}\|\bar{f})$	English	$\phi(\bar{e}\|\bar{f})$
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

Source: Philipp Koehn
http://www.statmt.org/book/slides/05-phrase-based-models.pdf

Responses from Translation Studies?

It is usually enough for translators who want to use [Google Translate] for initial drafting to know *nothing at all* about SMT

Robinson 2012

statistical-based MT, along with its many hybrids, is destined to turn most translators into posteditors one day, perhaps soon

Pym 2012

the question, in the long term, will not be *whether* translation will be done from the MT baseline, but simply *when* (and for which types of text and into which languages).

Garcia 2010

The dangers of knowing nothing at all about SMT…

- Analyses that miss the point:

  ‘Google Translate doesn’t understand the passive voice’

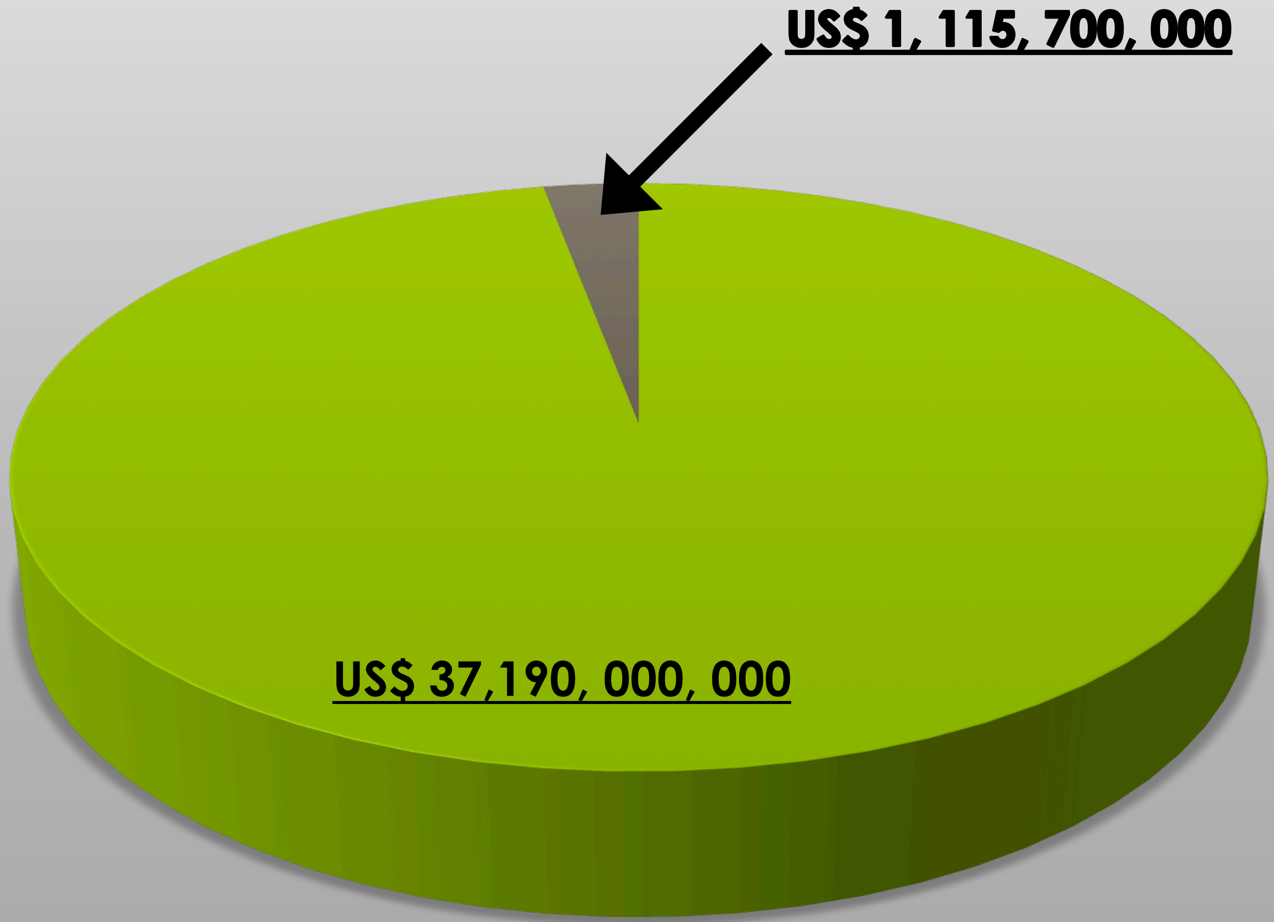
  ‘The SMT system fails to resolve structural ambiguities’

- Translators who miss opportunities:

  to learn, take ownership, capitalize on their own assets, be involved in the whole SMT workflow

The inevitability of post-editing?

Share of the translation services market accounted for by PEMT



Source: Common Sense Advisory 2014

Post-edited machine translation (PEMT)

'The contribution of PEMT to the overall market has been slowly creeping up over the last few years, but it is still relatively small ...'

Common Sense Advisory 2014

CSA 2012: 38.63%of LSPs offered PEMT

Almaghout *et al*. 2012: 42%; Doherty *et al*. 2013: 34%

And for which … and into which languages?

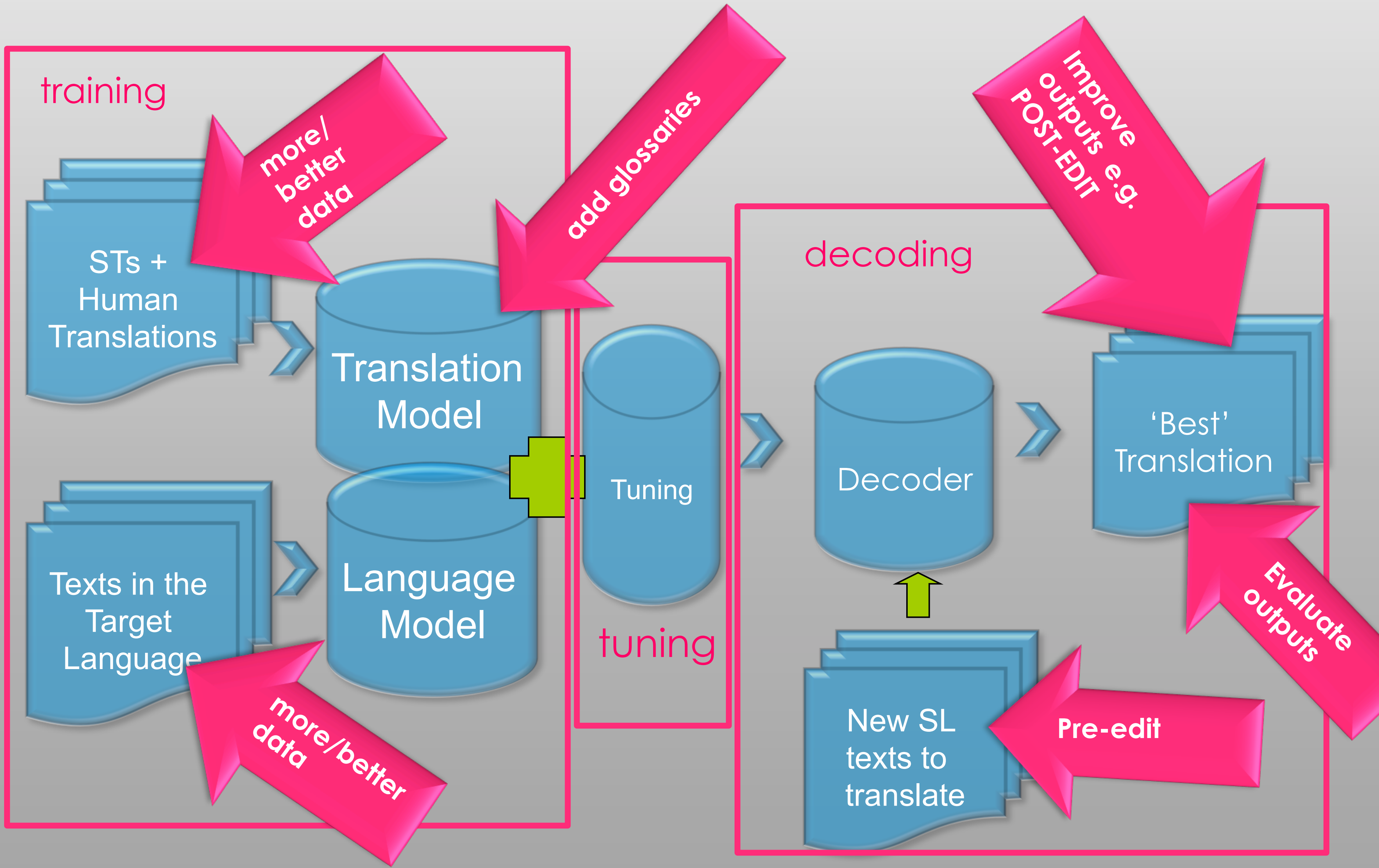
Source Language	Target Language										
	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	**31.2**	**32.1**	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	**30.1**	**31.1**	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	**30.5**	-	**40.2**	12.5	**32.3**	21.4	**35.9**	23.9
fr	23.7	18.5	26.1	**30.0**	**38.4**	-	12.6	**32.4**	21.1	**35.3**	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	**34.0**	**36.0**	11.0	-	20.0	**31.2**	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	**30.1**	**37.9**	**39.0**	11.9	**32.0**	20.2	-	21.9
sv	**30.3**	18.9	22.8	**30.2**	28.6	29.7	15.3	23.9	21.9	25.9	-

Table 2: BLEU scores for the 110 translation systems trained on the Europarl corpus

Source: Koehn 2005

And why just post-editing?

Possible Interventions to Improve MT Output



- training the **language model** on translated texts (rather than originals) in the Target Language improves system performance

Lembersky *et al*. 2012

- training the **translation model** on parallel texts that respect the translation task direction always gives the best results

Lembersky *et al*. 2013

- explained using insights from corpus-based translation studies, esp. Laviosa 1998

Possible Ways to *Evaluate* the Quality of MT Output

- Human Evaluation
- Automatic Evaluation

Adequacy

How much of the meaning is expressed?

5	All of the meaning
4	Most of the meaning
3	Much of the meaning
2	Little of the meaning
1	None of the meaning

as measured by LDC 2005 and in most shared-task MT evaluations

Fluency

How fluent is the translation?

5	Flawless English
4	Good English
3	Non-native English
2	Disfluent English
1	Incomprehensible

as measured by LDC 2005 and in most shared-task MT evaluations

MT Human Evaluation

KantanMT Language Quality Framework v 1.4

						Style						Syntax and Grammar					Tech		
Source	MT Target	Translator Improvement	Adequacy (Score 1-5)	Fluency (Score 1-5)	Overall quality (1-4)	Wrong terminology	Wrong Spelling	Source not Translated/Omissions	Compliance with client specs	Literal translation	Text/Information added	Capitalization	Wrong Word Form	Wrong Part of Speech	Punctuation	Sentence Structure	Tags and Markup	Locale Adaptation	Spacing
Enter source text in this column	*Enter Machine Translated texts in this column*	*Enter Translator's improved translation in this column*																	

Automatic Evaluation Metrics

Basic strategy:

- Compare MT output or 'candidate' with a 'gold standard' or 'reference' translation produced by a human.

Example:

Source: *Ceci n'est pas une pipe.*

Reference: This is **not a pipe**.

MT candidate: That was **not a pipe**.

Automatic Evaluation Metrics

F-measure

Asks:

- How many words did the system get right? (precision)
- Did it fail to produce some good words? (recall)

Automatic Evaluation Metrics

Translation Error Rate

Asks:

How many edits would you have to make to get from the machine output to the human reference?

Automatic Evaluation Metrics

BLEU

Asks:
How many *n*-word sequences (*n*-grams) in the candidate translation also occur in the reference translation?

Example:
This is **not a pipe**.
This was definitely **not a pipe**.
share one 1-gram and one 3-gram…

t Mail Server Web Ac | Kantan Machine Translation | www.kantanmt.com : List all | Dublin City University - Caler

www.kantanmt.com/jobs.php?&page=1

KantanMT.com

Dorothy Kenny | Logout | Signup

Home Company Our Platform My KantanMT

Track Your Jobs

Use this screen to track the progress of all your KantanMT jobs.

Click to toggle view of KantanMT usage.

I can't create a graph without at least 3 site activities! Why not improve your KantanMT engine by uploading more training data!

Display 15 records Search:

#	Status	Engine	Job Name	Started
11617	KantanMT [OK]	Default-Profile	Building KantanMT Engine for Default-Profile	Feb 28th, 10:35
11618	Complete...	Default-Profile	Quick Launch for Default-Profile	Feb 28th, 10:58

BLEU Score

F-Measure Score

TER Score

Disk Space Usage

Word Count Statistics

11%

217.1K

This is the number of words in your training data as a percentage of 2m words.

Practical SMT Assignment at DCU

5th iteration in autumn 2015

- Source data to train an SMT engine
  e.g. DGT Translation Memory (24 languages, TMX format)
- Profile the data
  languages, domain, character encoding, irrelevant/ corrupt data, number of one-to-many translations, number of unique words, etc.
- Upload data to SMT system (KantanMT.com) and train own engine

SMT Practical Assignment continued

- Test the engine with new input
  - in-domain *and*
  - out-of-domain
- Evaluate the output using human metrics and AEMs
- Devise a strategy to improve system performance

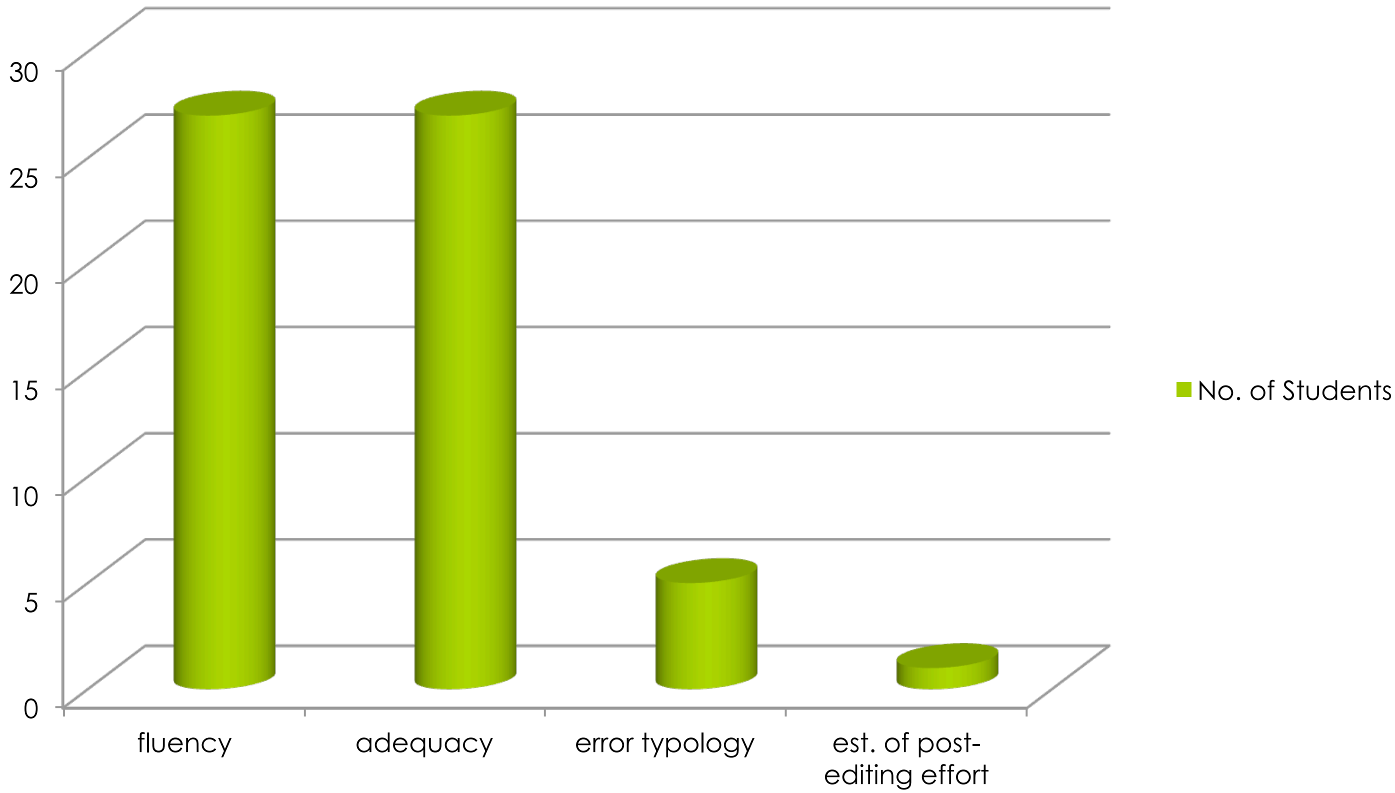
  Pre-edit input files?

  Create and integrate a glossary?

  Add more data?

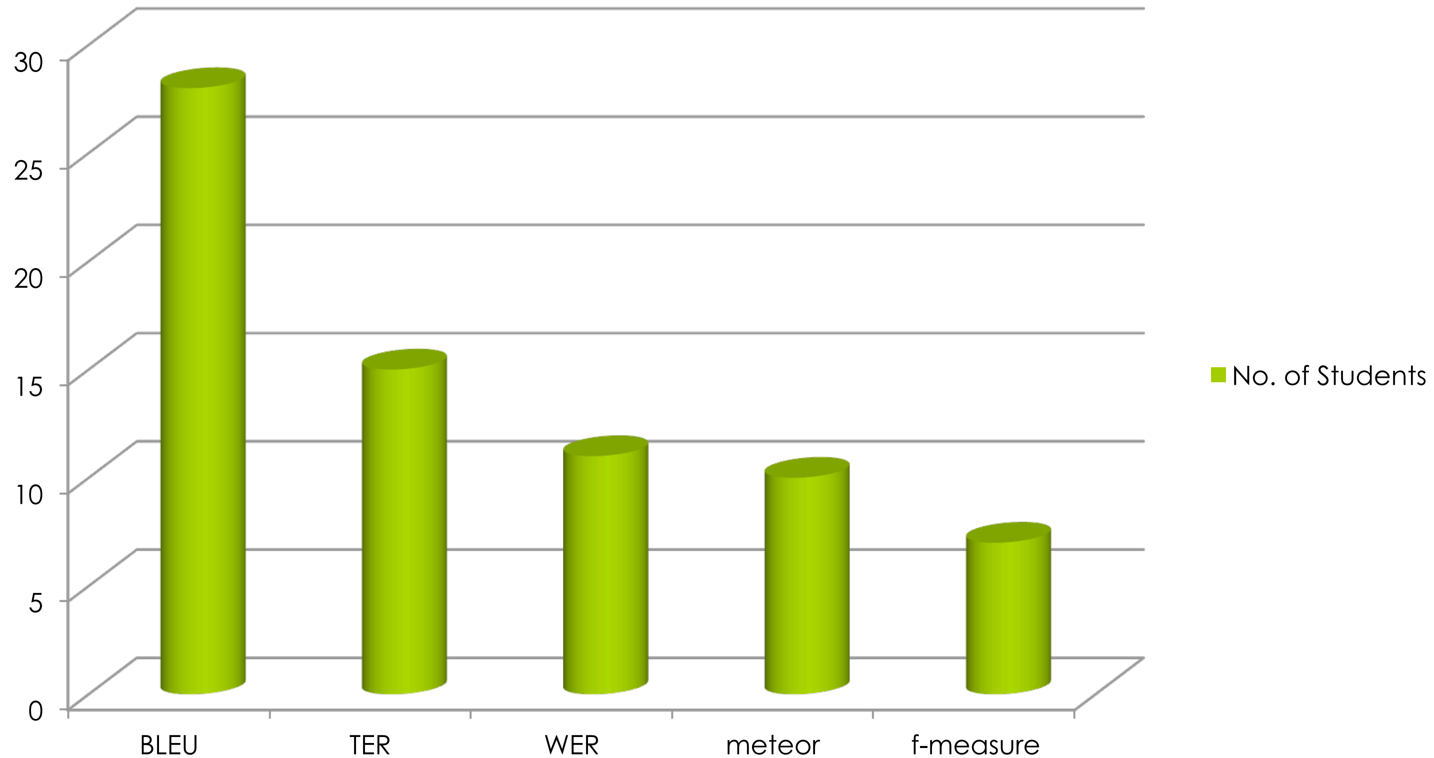
  Post-edit the output?
- Apply input-based strategies and test system again

Evaluating SMT output using human metrics



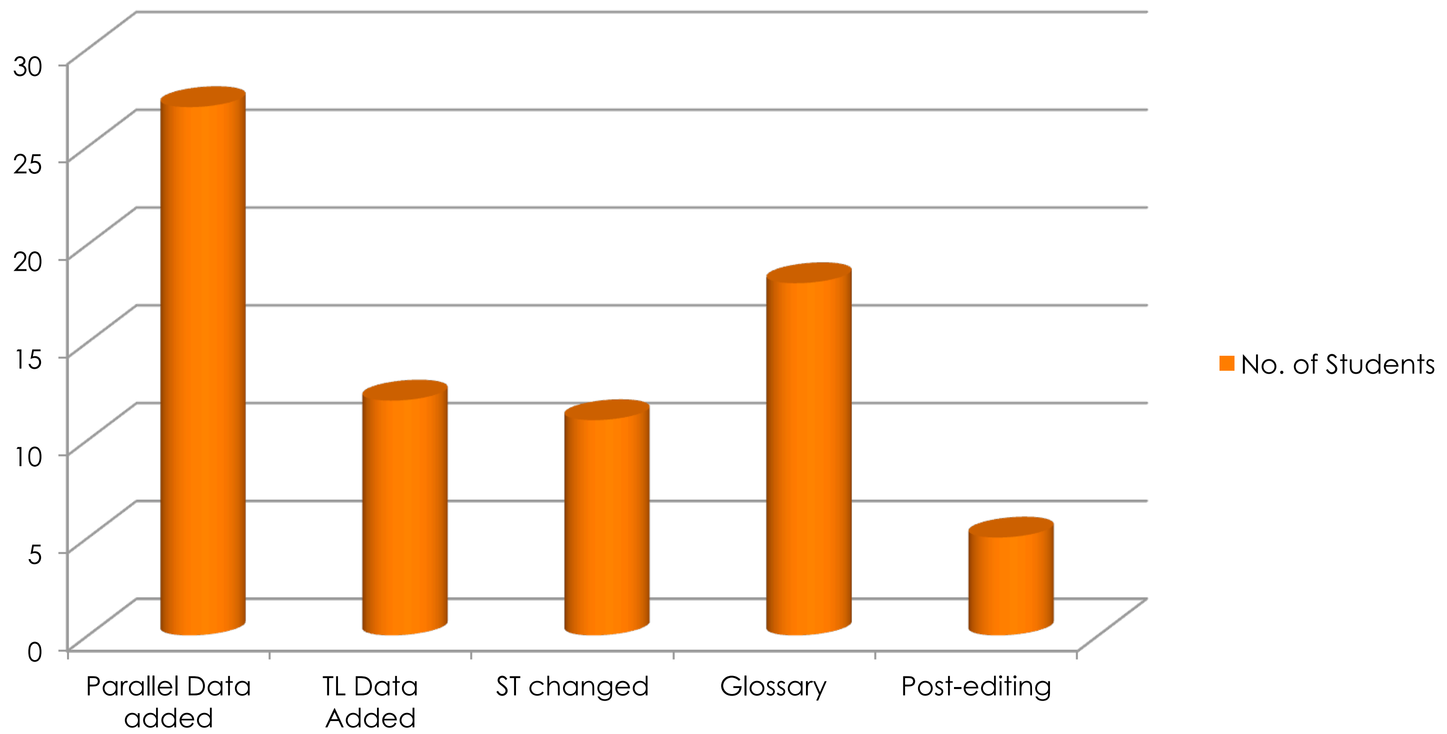
Human evaluation metrics used by students
29 students; 2 metrics each on average (2013 data)

Evaluating SMT output using AEMs



Automatic Evaluation Metrics (AEMs) favoured by students
29 students; multiple AEMs used by all. (2013 data)

Interventions to improve SMT output



29 students; 2.5 interventions each on average (2013 data)

Sample Student Conclusions (1)

‘This research has shown that, give a full domain-specific corpus and *in-domain* texts to translate, KantanMT produces translations of a very good quality and that for in-domain texts, the Bleu and TER ratings for the actual engine are a good indicator of the quality of its output. MT output can be improved by avoiding long, convoluted sentences and by the use of glossaries when translating new areas within the domain.’

Sample Student Conclusions (2)

- “The research demonstrated that the machine translatablity of out-of-domain source texts can be increased, with a glossary having the most impact.”
- Student wonders whether it was worth the effort put into the out-of-domain glossary however and concludes:

  “Expert advice to stick to one engine for one topic springs to mind here.”

Employment?

Three of DCU's MA in Translaton Studies/MSc in Translation Technology graduates now working with this SMT provider

- in positions that didn't exist 3 years ago
- in a company that didn't exist 3 years ago

Professional ethics and SMT

In the case of Free Online MT

- breach of contract
- confidentiality (despite ‘anonymisation’)
- permission to use others’ work
- attribution (Drugan and Babych 2010)

Even with in-house uses of SMT

- quality assurance (Karamanis *et al*. 2011)

Translators constantly concerned about having to work to lower quality standards, to produce just ‘good enough’ (post-edited) translation (see Drugan 2013)

Translation ethics (Chesterman 2001)

Typical foci:

- representation of STs, authors, the Other
- service, communication, norms

What about the world's responsibility towards translators?

This would belong to 'a general ethics of translation and translatorial behaviour'.

Translation ethics (Chesterman 2001 cont.)

virtue: ‘an acquired human quality that helps a person strive for excellence in a practice’

in order to make the best ethical decisions, the most important virtue that a translator can possess is the desire to make the right decision:

‘the translator must *want* to be a *good* translator, must strive for excellence in the practice of translation’

The post-editor's dilemma?

How can you strive to be your best at being just 'good enough'?

Ethics and SMT
Translation as a natural resource

'[Microsoft's] largest available natural resource is the nearly two decades of product documentation that has been localized into an increasing range of languages and preserved as translation memories.'

Joscelyne (2009)

Natural Resource Metaphor

~45 instances of discovery of/discover*+dir obj

N	Concordance
17	exists. However, this was not sufficiently balanced by a conservative discovery of correspondences in the pair extraction algorithm. Although the
18	subset of the phrase pairs in the Baseline+Syntax model.10 This discovery is a very positive and interesting by-product of these experiments.
19	cognates (Sect. 4.1) and an algorithm for applying the induced rules to the discovery of potential cognates in a corpus (Sect. 4.2). 4.1 Learning
20	and used to group together related data. This approach supports the discovery of knowledge from the acquired data (i.e. from the ground) instead
21	Till now, the focus of most of the investigations in this field has been on the discovery and pairing of bilingual sites, domains, HTML documents and
22	available texts, we propose and investigate a general framework for the discovery of such expressions from comparable corpora. The main practical
23	not received much attention, possibly because in many applications the discovery of translationally equivalent vocabulary items was the main goal
24	to search for their translations in the French sentences. This makes the discovery of the different translation possibilities more difficult (adaptées à,
25	text needs to be produced by translators before it can be used for the discovery of new dictionary entries. Bilingual comparable corpora are an

Natural Resource Metaphor

9 instances of MINE in the *Machine Translation* Corpus, e.g.:

N	Concordance
1	parallel alignment approaches. We are interested in (1) how to mine web content and prepare HTML files for bilingual text alignm
2	of bilingual corpus building. What we need is to seek ways to mine the widely available web resources to bridge the gap in the
3	accessing server-resident databases. Translators can still mine their own databases, and expand them with the translation u
4	previously have been shown to profit from translation equivalents mined from comparable corpora, including construction of probabi
5	translation extraction of web-based materials is mostly related to mining web contents as a bilingual corpus and aligning the bilingu
6	on data source acquisition for natural language processing tasks by mining the world wide web. Representative works related to the

655 instances of EXTRACT in the *Machine Translation* Corpus, e.g.:

N	Concordance
370	texts. These are also the two primary challenges facing systems that extract translations from bilingual websites. As we know, the building
371	rder are going to be preferable to orderings which are very different. We first extract permutations from alignments, and then apply standard dista
372	parser with deterministic head-finding, while Owczarzak et al. (2007a) extract the semantic dependency relations from an LFG parser (Cahi
373	to lemmas and suffixes. 4. Apply morphological transformation rules. 5. Extract the surface string. For syntactic transformation, we propose a
374	original is aligned with its translation at the sentence level) can be used to extract newdictionary entries with high accuracy (Dagan and Church
375	nodes. Starting from this definition, a linear-time algorithm is proposed to extract translation rules through one-time traversal of the leaf nodes i
376	make use of new tools to automatically build a large parallel treebank and extract a set of linguistically-motivated phrase pairs from it. We show
377	news releases in the StatCan publication The Daily. The goal is to extract translations for translation memory systems, for translation
378	Abstract Statistical methods to extract translational equivalents from non-parallel corpora hold the pro

Metaphors in SMT

We have simply observed **conventional metaphors** used with any kind of data/information.

But conventional metaphors are particularly potent carriers of latent ideology…

natural resource:

translation as an undifferentiated mass, a resource accessible to those with the technology/legal rights to exploit it

Ethics and Translation Teaching

If we move straight to 'translating from MT',

what other previous, hard-earned accomplishment might be put in jeopardy?

(Garcia 2010, 2011)

(Hirshmann 1991; Morozov 2013)

Moving towards a conclusion

Technoneutrals

“on the one hand this, but on the other hand that”

(Tehranian 1990, Morozov 2013)

Technostructuralism

The impact of a technology:

- is always mediated through institutions and social forces
- does not flow from the inherent characteristics of the technology
- is not neutral
- depends on the context

Technostructuralism

Technology is studied by

"analyzing how particular aspects of a given technology … might restructure political and social relations, introducing entirely new classes of actors into the game"

Morozov (2013: 171)

Shifting identities

- ‘cross-language carriers’ (TAUS 2007)
- ‘post-editor’ used to describe person editing matches from translation memory, as well as translations from an MT engine (eg Guerberof 2008, He 2011)
- ‘monolingual translator’ used to describe post-editor with no knowledge of the SL (Koehn 2010)
- ‘translating by post-editing’ (Garcia 2011)
- crowdsourcing and the re-positioning of translation as non-work, done by ‘friends’ (Google Translator Toolkit video)

Boundary Shifters

The actors in my study also morph their identities. Musicians on occasion turn into salesmen; engineers on occasion turn into musicians; and engineers can become salesmen… When the modular Moog synthesizer was first used in recording studios, no one knew what to call its operators: were they engineers, programmers, producers, musicians, or what?...I call these actors "boundary shifters.'

Pinch (2008)

Holistic deployment of SMT in translator training

(we hope) promotes

- sensitivity to context (language pairs, text types, training data, etc.)
- the adoption of new roles/introduction of new actors
- appreciation of the role of institutions (companies, the EU, universities)
- student awareness of their own agency

Conclusions

SMT brings with it great opportunities.

The time is ripe for even more research into translation technology that comes from within the humanities and social sciences.

We need more sociologically-oriented, ethically-aware, critical research in translation technology.

This is as true for translator training as it is for any other branch of translation studies.

Bring it on!

Thank you!

Some References

Doherty, Stephen and Dorothy Kenny. 2014. The Design and Evaluation of a Statistical Machine Translation Syllabus for Translation Students. *The Interpreter And Translator Trainer*, 8, 2, pp295-315.

Kenny, Dorothy and Stephen Doherty. 2014. Statistical Machine Translation in the Translation Curriculum: overcoming obstacles and empowering translators. *The Interpreter And Translator Trainer*, 8, 2, pp276-294.

Kenny, Dorothy. 2011. *The ethics of machine translation*. In: New Zealand Society of Translators and Interpreters Annual Conference 2011, 4-5 June 2011, Auckland, New Zealand.