FACULTY OF ARTS AND PHILOSOPHY

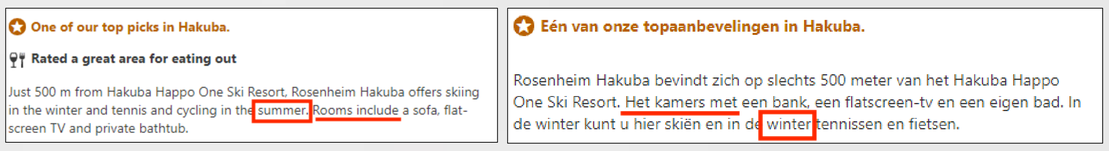
LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

Lieve Macken, Joke Daems & Arda Tezcan

ARISTOCAT: ASSESSING THE COMPREHENSIBILITY OF AUTOMATIC TRANSLATIONS

Project goals

- Readers are more often confronted with 'raw' (unedited) MT output due to increased use of MT
- But MT systems cannot guarantee that the text they produce is fluent and coherent in both syntax and semantics, leaving the reader to guess parts of the intended message
- How do end users engage with raw machine-translated text?



- Assess comprehensibility of automatic translations
- Collect and analyse eye movements of participants reading Dutch machine-translated text
- Investigate the impact of different categories of MT errors on comprehension
- Automatically predict the MT errors that hamper comprehension most in Dutch machine-translated text

How to assess comprehension?

- 3 texts selected from the English MT Evaluation version of CREG (CREG-MT-eval)
- 3 Dutch translations for each text: DeepL, GNMT, HT
- 99 participants (each participant read 2 different translated texts: HT-MT or MT-MT)
- 5 reading comprehension questions per text + overall clarity score 1-5

Averaged clarity score	Text 1	Text 2	Text 3
Human Translation	**4.1**	**4.1**	**4.0**
Google Translate	3.5	3.5	3.1
DeepL	3.2	3.4	3.5

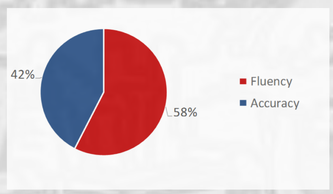
Averaged comprehension score	Text 1	Text 2	Text 3
Human Translation	**3.4**	2.4	3.1
Google Translate	3.0	1.6	3.3
DeepL	2.4	**2.6**	**3.5**

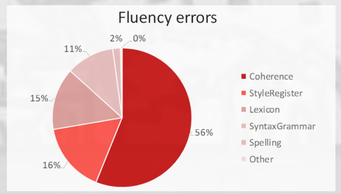
- HT best clarity scores, but large variation across participants
- Incongruent results: HT best overall clarity scores ↔ DeepL best comprehension scores for 2 texts
- Clarity scores and reading comprehension test assess different aspects of reading comprehension?

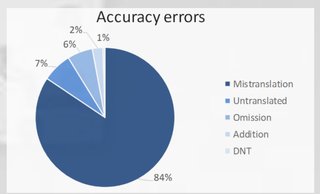
➢ Macken & Ghyselen (2018). Measuring comprehension and perception of neural machine translated texts : a pilot study (Proceedings of TC40)

MT for literary translation?

- Challenges: fragmented views of context, figurative language, cultural references, lexical richness …
- Agatha Christie's novel *The Mysterious Affair at Styles* (Google Translate – May 2019)
- Assess NMT quality on literary texts in Dutch (first chapter, 4358 words)



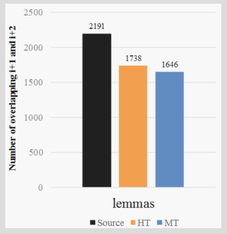




- Compare lexical richness and local cohesion in NMT output and HT (whole novel, 56000 words)
- Type-token ratio + variants (sensitive to text length), mass index and mean segmental TTR
- Lexical overlap between a given sentence and the succeeding sentence(s)

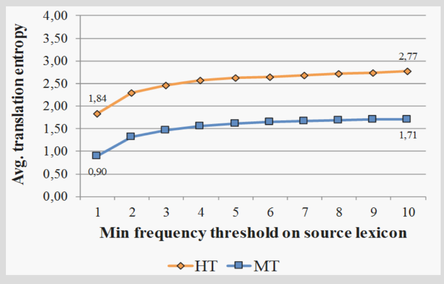
Lexical richness	Source	HT	MT
TTR	0.073	0.079	**0.083**
Root TTR	19.71	21.56	**22.17**
Corr. TTR	13.94	15.24	**15.68**
Mass index	**0.021**	0.020	0.020
Mean segmental TTR	0.648	**0.670**	0.660

- Local lexical cohesion
- HT contains more overlapping lemmas of content words than MT



- (Average) word translation entropy = degree of uncertainty to choose a correct translation from a set of target words

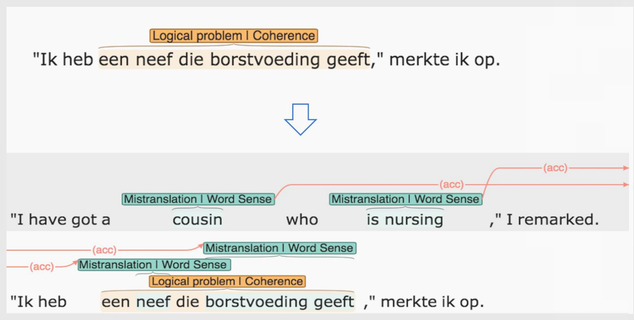
Source	MT (prob.)	HT (prob.)
funny	grappige (0.57)	grappig (0.22)
	grappig (0.29)	grapjas (0.22)
	grappigs (0.14)	leuk (0.22)
		gekke (0.22)
		wel (0.11)
WTE	= 1.37	= 2.27



➢ Tezcan, Daems, & Macken (2019). When a`sport' is a person and other issues for NMT of novels (Proceedings of the Qualities of Literary Machine Translation)

Quality of MT output?

- Two-step approach for error annotation
- Fluency + accuracy (WebAnno)
- Corpus of 665 sentences (< DPC)
- RBMT (Systran)
- SMT (Google Translate, June 2014)
- NMT (Google Translate, June 2017)



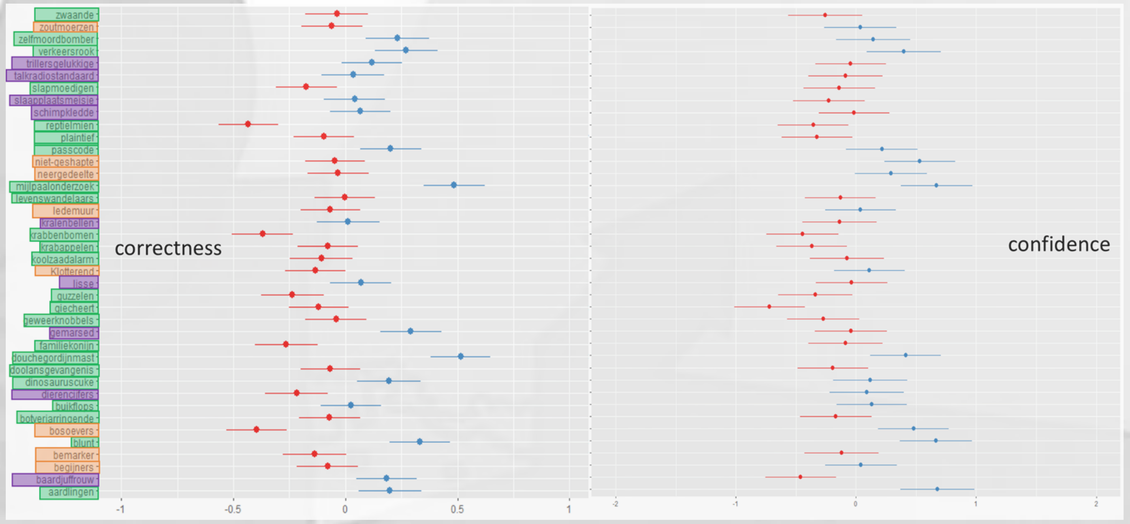
Fluency Errors	RBMT	SMT	NMT
Grammar	863	936	255
Orthography	280	244	94
Lexicon	535	232	365
Multiple errors	144	112	7
Other	0	1	0
Total	**1823**	**1525**	**721**

Accuracy errors	RBMT	SMT	NMT
Mistranslation	970	477	319
DNT	116	14	23
Untranslated	66	67	48
Addition	60	41	1
Omission	43	115	62
Mechanical	52	20	11
Total	**1307**	**734**	**464**

➢ Van Brussel, Tezcan & Macken (2018). A fine-grained error analysis of NMT, PBMT and RBMT output for English-to-Dutch (Proceedings of LREC)

How comprehensible are newly invented words in NMT output?

- NMT operates at sub-word level to reduce vocabulary size and can 'invent' new words, e.g. *bekinnen* as translation for *pelvic fins* (*pelvic = bekken + fins = vinnen*) or *familiekonijn* as translation for *family rabbi*
- 86 participants were given 15 non-existing words (5 single words; 10 compounds)
- Describe the meaning or select the correct meaning from a predefined list in two conditions: words in isolation vs. in sentence context + participants had to indicate confidence
- 60% wrong answers; sentence context had a positive impact on correctness and confidence



➢ Macken, Van Brussel & Daems (submitted) NMT's wonderland where people turn into rabbits. A study on the comprehensibility of newly invented words in NMT output (CLIN Journal)

➢ Macken (2019) Mysterie van de dag: waarom vindt een automatisch vertaalsysteem soms nieuwe woorden uit? Knack.

Future work

- MT Error annotations on whole novel
- Extend Ghent Eye-Tracking Corpus (GECO) with MT version
- Compare reading behaviour HT vs NMT
- Analyse impact of different types of MT errors on reading behaviour
- Build ML system to predict comprehensibility of machine-translated text/sentences

ArisToCAT is a four-year research project funded by the Research Foundation - Flanders (FWO) – grant number G.0064.17N

https://research.flw.ugent.be/projects/aristocat

Contact:

- Prof. Dr. Lieve Macken – lieve.macken@ugent.be
- Dr. Joke Daems - joke.daems@ugent.be
- Dr. Arda Tezcan - arda.tezcan@ugent.be

GHENT UNIVERSITY