RESEARCH QUESTIONS

- How does human translation differ from post-editing machine translations?
- How does translation revision differ from post-editing machine translations?
- How can we objectively assess translation quality?

- How can we measure translation difficulty?
- Can we automatically predict whether a text is difficult to translate

PRODUCT

```
The intense interest aroused in the public by what was known at the time as 'The

Styles Case' has now somewhat subsided. Nevertheless, in view of the world-wide

notoriety which attended it, I have been asked, both by my friend Poirot and the

family themselves, to write an account of the whole story. This, we trust, will

effectually silence the sensational rumours which still persist. I will therefore

briefly set down the circumstances which led to my being connected with the

affair.
```

GHENT
UNIVERSITY

PRODUCT

The intense interest aroused in the public by what was known at the time as 'The

Styles Case' has now somewhat subsided. Nevertheless, in view of the world-wide

notoriety which attended it, I have been asked, both by my friend Poirot and the

family themselves, to wr

effectually silence the

briefly set down the cir

affair.

De enorme belangstelling die het publiek toonde voor wat indertijd bekend stond

als 'de zaak Styles', wordt nu wat minder. Niettemin is mij, zowel door mijn

vriend Poirot als door de betrokken familieleden, verzocht een verslag van het

hele gebeuren te schrijven, gezien het feit dat er in de hele wereld grote

ruchtbaarheid aan gegeven is. Op deze manier hopen we voorgoed een einde te maken

aan de sensationele geruchten die nog steeds de ronde doen.

GHENT
UNIVERSITY

PRODUCT

- Is there a difference in quality between HT and PE?
- Is there a difference in the most common error types in HT and PE?
- Can readers tell whether a text was translated from scratch (HT) or post-edited MT?
- How does artificially generate language (MT) differ from human language? Can we “measure” this difference?

PROCESS

GHENT
UNIVERSITY

PROCESS

GHENT
UNIVERSITY

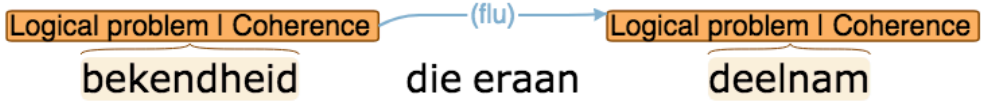
PROCESS

- Is PE faster than HT?
- Is PE cognitively more demanding than HT?
- Are more (or other) external resources consulted in HT compared to PE?
- How do translators interact with (the interface of) translation technology tools?
- What are the typical source text segments that pose problems for translation?
- Is there a difference between students and professional translators?

PRODUCT

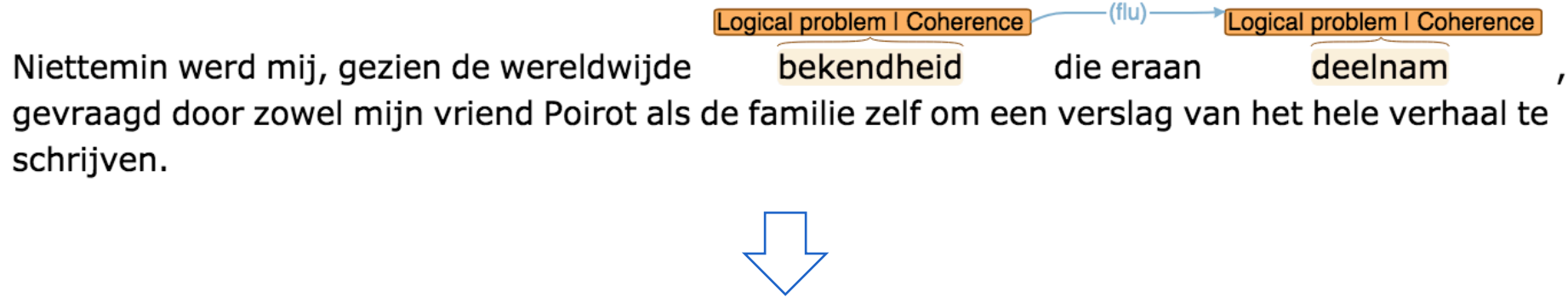
GHENT
UNIVERSITY

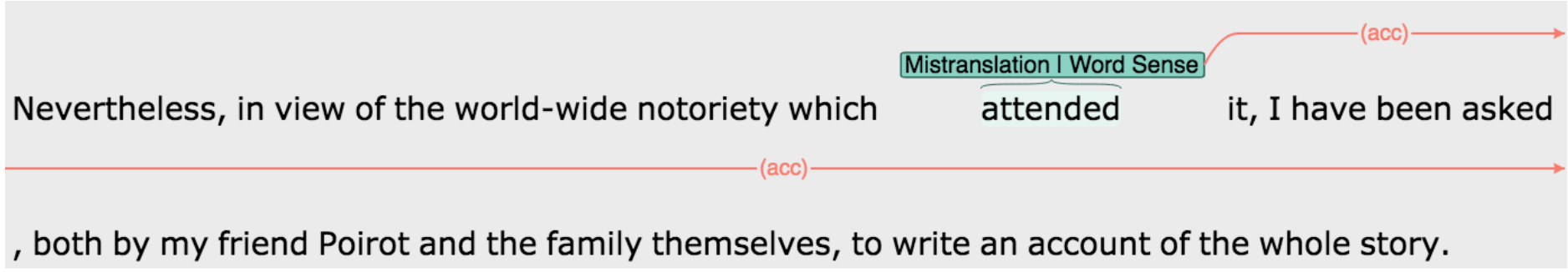
ERROR ANNOTATION (WEBANNO)

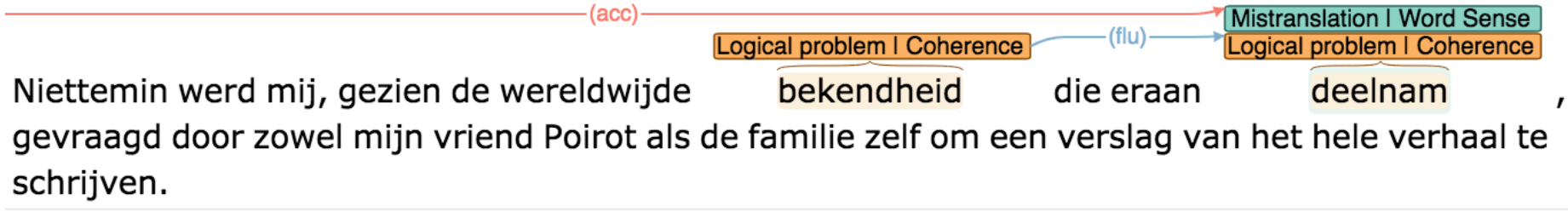


GHENT UNIVERSITY

ERROR ANNOTATION (WEBANNO)







MT QUALITY IMPROVEMENTS

TARGET

Fluency Errors	GT 2014	GT 2017
Grammar	936	255
Orthography	244	94
Lexicon	232	365
Multiple errors	112	7
Other	1	0
Total	**1525**	**721**

SOURCE & TARGET

Accuracy errors	GT 2014	GT 2017
Mistranslation	477	319
DNT	14	23
Untranslated	67	48
Addition	41	1
Omission	115	62
Mechanical	20	11
Total	**734**	**464**

GHENT UNIVERSITY

Van Brussel, Tezcan & Macken. 2018. "A Fine-grained Error Analysis of NMT, PBMT and RBMT Output for English-to-Dutch." LREC

HOW DO TRANSLATIONS DIFFER?

- Translation edit rate
- Lexical richness
- Cohesion
- Syntactic equivalence

} Linguistic characteristics

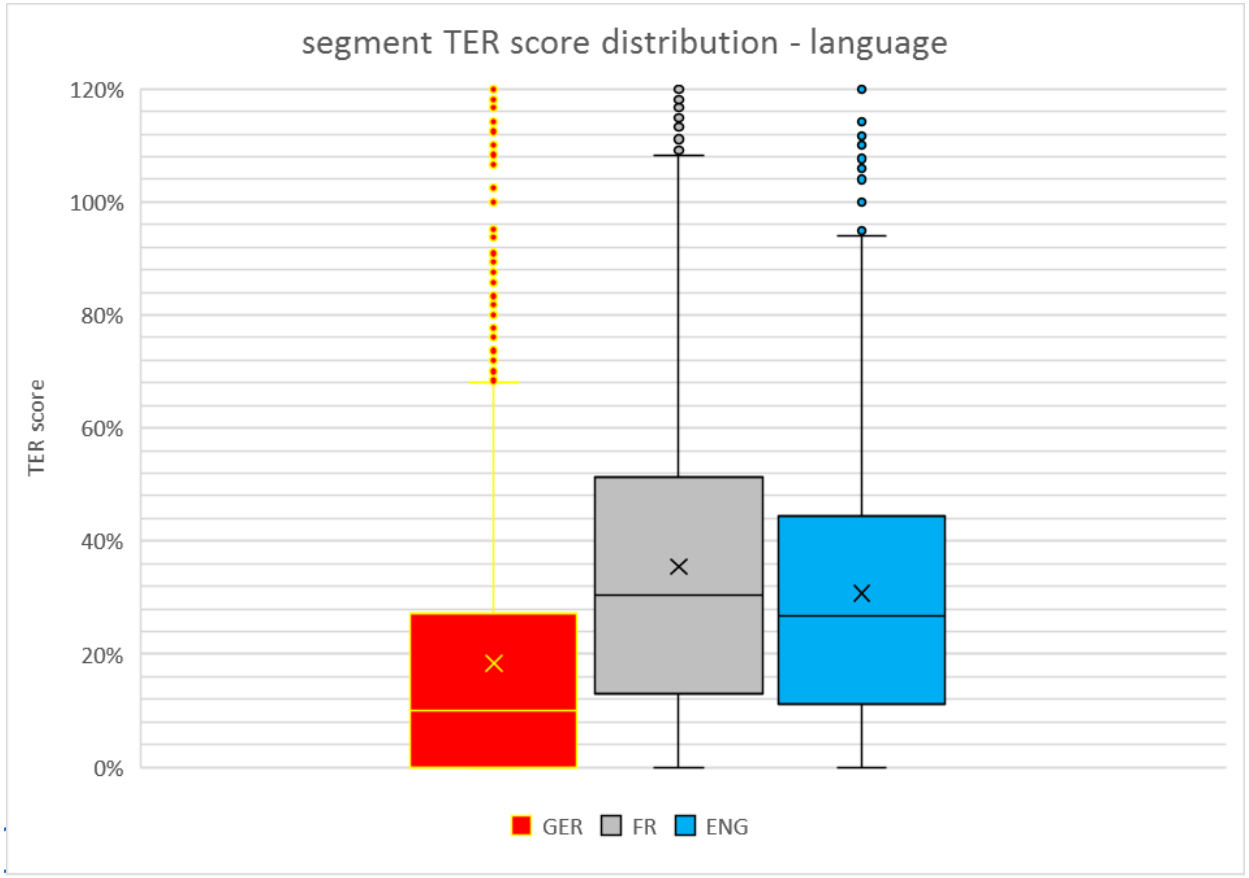
TRANSLATION EDIT RATE (TER)

```
_  _           _    Not only  Notre Dame 's works of  art were saved
In addition to  the saved Notre Dame _  works _   _   _    _

, but also the bees of the cathedral _     _     survived the fire .
, _   _    the bees of the cathedral have also survived the fire .
```

HTER = 11/22 (0.5)

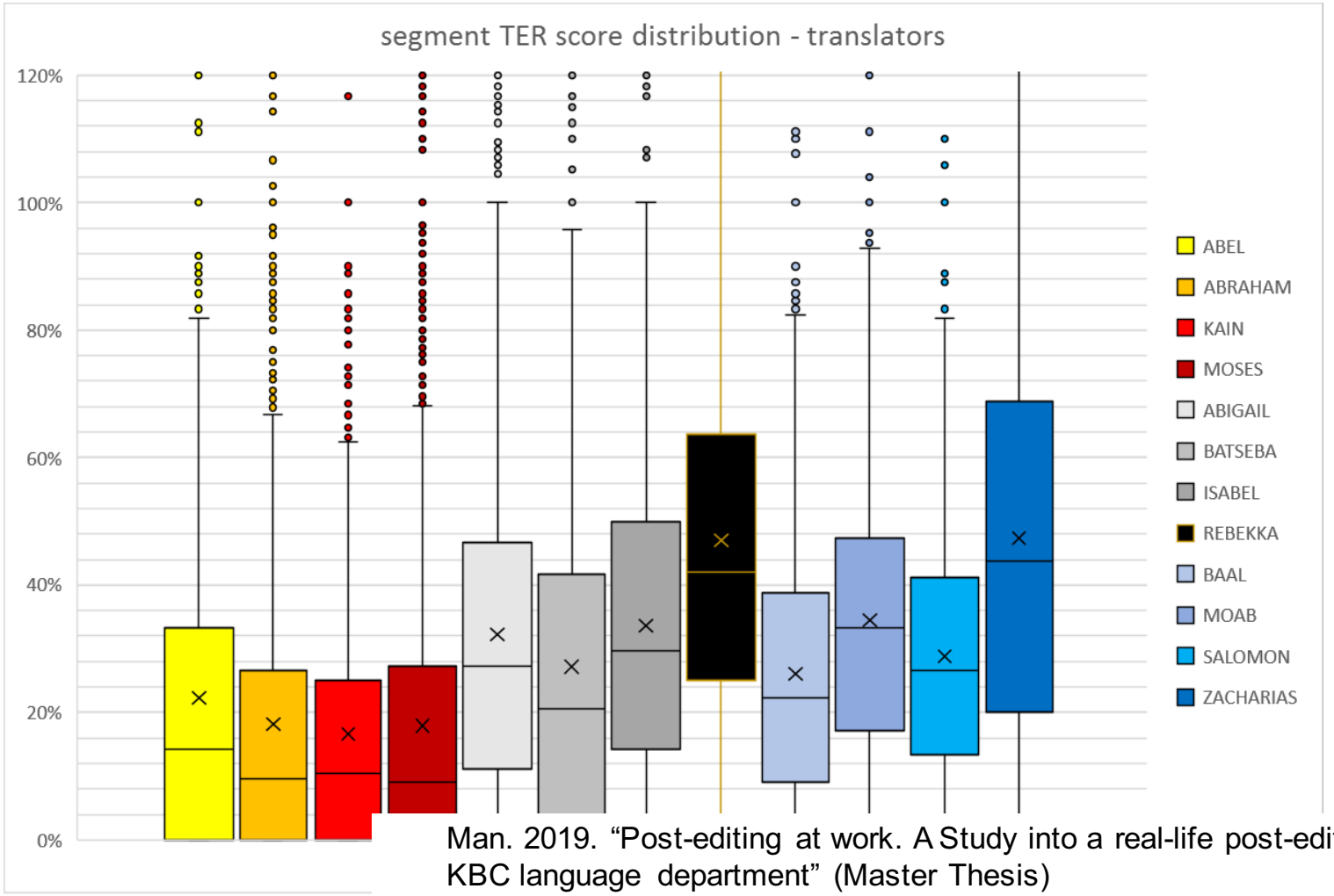
TER SCORE DISTRIBUTIONS: LANGUAGES



GHENT UNIVERSITY

Man. 2019. “Post-editing at work. A Study into a real-life post-editing environment at the KBC language department” (Master Thesis)

TER SCORE DISTRIBUTIONS: TRANSLATORS



Man. 2019. “Post-editing at work. A Study into a real-life post-editing environment at the KBC language department” (Master Thesis)

LEXICAL RICHNESS

- Type-token ratio → No. unique words
- Mean Segmental TTR → Average TTR on subsets of 100 words

	ST	HT	MT
TTR	0.073	0.079	0.083
MSTTR	0.648	0.670	0.660

- Inconclusive results

GHENT UNIVERSITY

LEXICAL RICHNESS

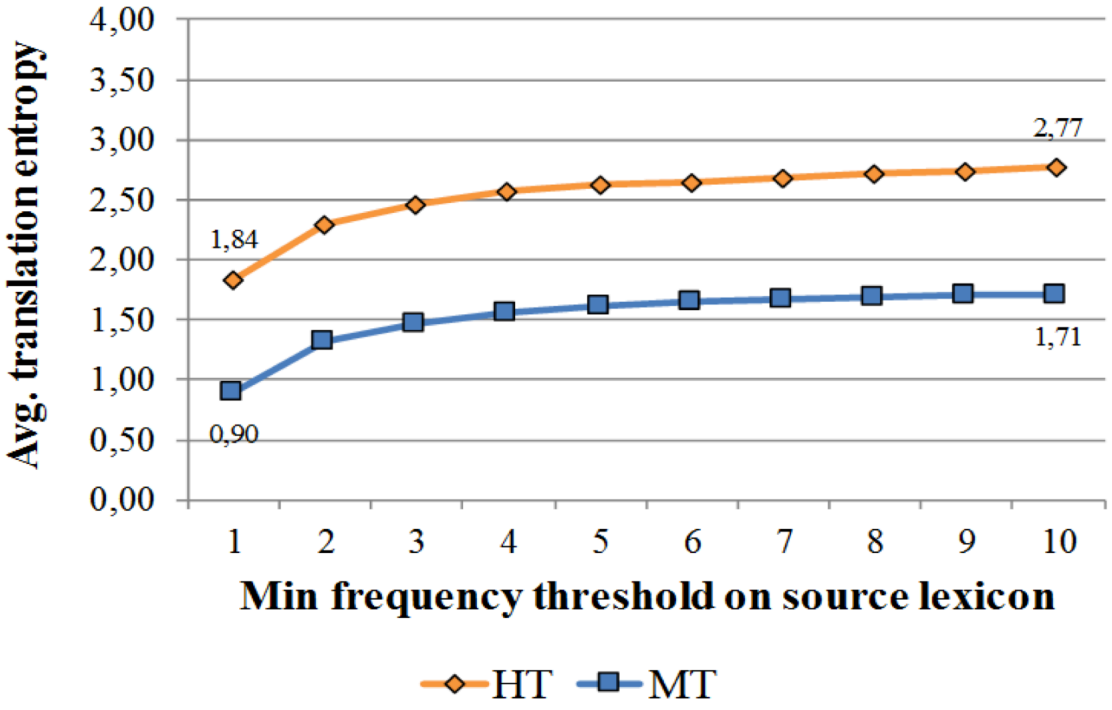
- Word Translation Entropy

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

GHENT UNIVERSITY

LEXICAL RICHNESS

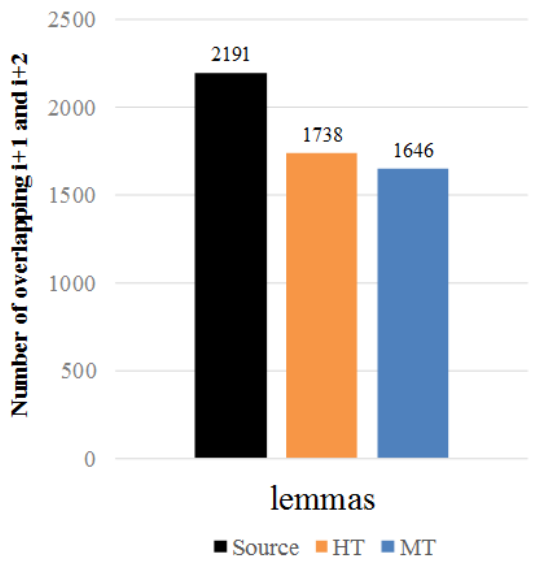
- Word Translation Entropy



Tezcan, Daems, & Macken (2019). When a `sport' is a person and other issues for NMT of novels

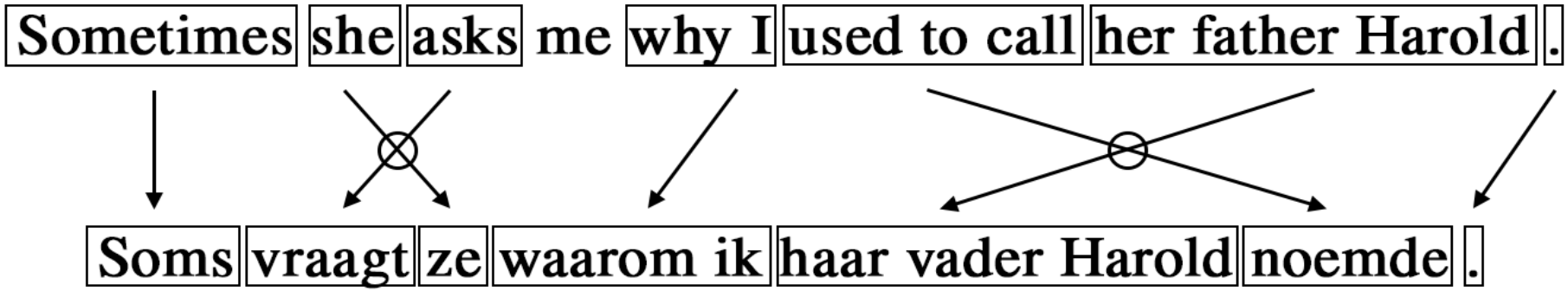
COHESION (TWO SUCCESSIVE SENTENCES)

- Lexical cohesion: overlapping lemmas of content words (nouns, verbs, adjectives and adverbs)
- Semantic cohesion: overlapping synonyms of lemmas of content words



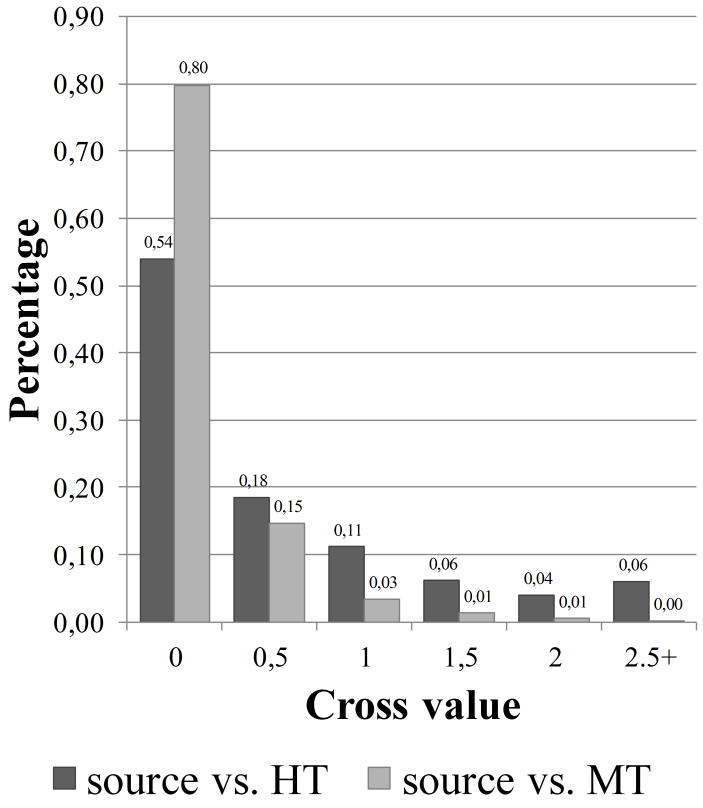
SYNTACTIC EQUIVALENCE

- Amount of re-ordering



SYNTACTIC EQUIVALENCE

- 80% of MT sentences have low cross value
- MT follows structure of ST more closely than HT



Tezcan, Daems, & Macken (2019). When a `sport' is a person and other issues for NMT of novels

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)

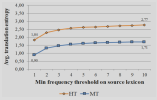
- Compare lexical richness and local cohesion in NMT output and HT (whole novel, 56000 words)
- Type-token ratio + variants (sensitive to text length), mass index and mean segmental TTR
- Lexical overlap between a given sentence and the succeeding sentence(s)

Lexical richness	Source	HT	MT
TTR	0.075	0.079	**0.085**
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 (0.13)
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)

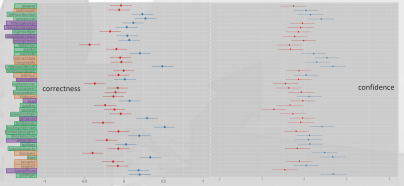
Fluency Errors	RBMT	SMT	NMT
Grammar	863	936	255
Orthography	280	244	94
Lexicon	535	232	365
Multiple errors	144	112	7
Other	0	1	0
Total	**1823**	**1525**	**721**

Accuracy errors	RBMT	SMT	NMT
Mistranslation	970	477	379
DNT	116	14	25
Untranslated	66	67	48
Addition	60	41	1
Omission	43	115	62
Mechanical	52	20	11
Total	**1307**	**754**	**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

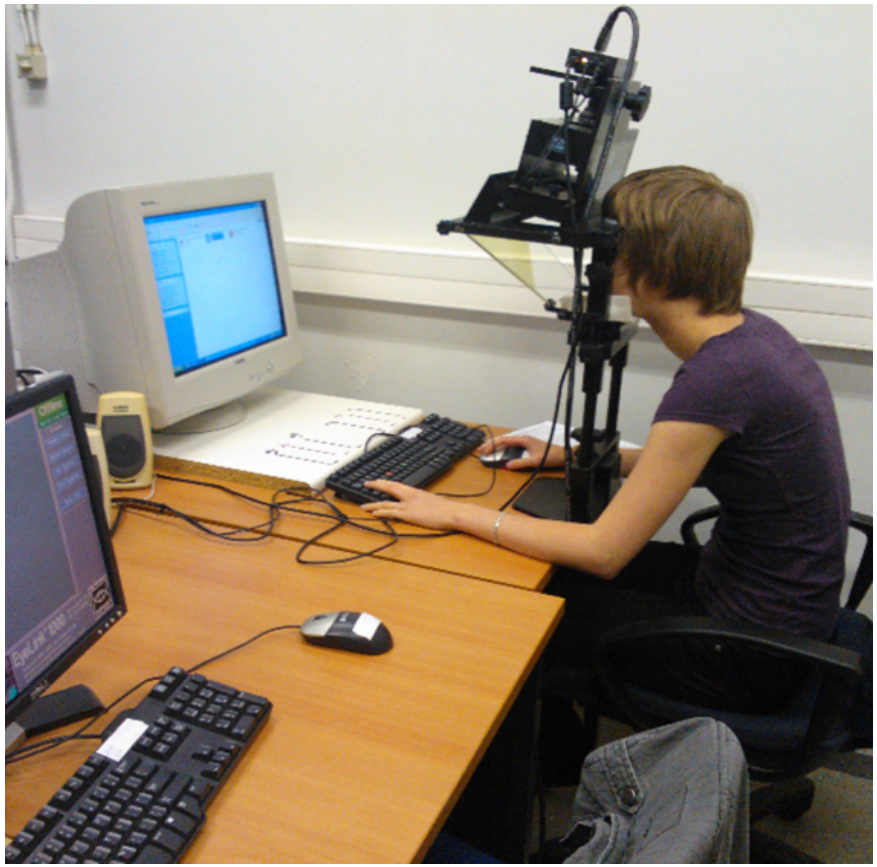
GHENT UNIVERSITY

PROCESS

GHENT
UNIVERSITY

PROCESS

- UAD
- Keystroke logging
- Eye-tracking
- Screen capture



GHENT UNIVERSITY

KEYSTROKE LOGGING

Registers all keystrokes & mouse movements

#Id	Event Type	Output	Position	DocLength	Character Production	StartTime	StartClock	EndTime	EndClock	ActionTime	PauseTime	PauseLocation
0	focus	Wordlog.docx - Microsoft Word			0	4259	00:00:04	4259	00:00:04	0	4259	UNKNOWN PAUSE
3	keyboard	D	0	1	1	8050	00:00:08	8175	00:00:08	234	7941	INITIAL PAUSE
4	keyboard	e	1	2	2	8455	00:00:08	8580	00:00:08	125	405	WITHIN WORDS
5	keyboard	m	2	3	3	8767	00:00:08	8861	00:00:08	94	312	WITHIN WORDS
6	keyboard	o	3	4	4	8986	00:00:08	9157	00:00:09	171	219	WITHIN WORDS
7	keyboard	n	4	5	5	9079	00:00:09	9251	00:00:09	172	93	WITHIN WORDS
8	keyboard	s	5	6	6	9329	00:00:09	9516	00:00:09	187	250	WITHIN WORDS
9	keyboard	t	6	7	7	9625	00:00:09	9688	00:00:09	63	296	WITHIN WORDS
10	keyboard	r	7	8	8	9797	00:00:09	9891	00:00:09	94	172	WITHIN WORDS
11	keyboard	a	8	9	9	9891	00:00:09	10047	00:00:10	156	94	WITHIN WORDS

(source: http://www.inputlog.net/wp-content/uploads/Inputlog_manual.pdf)

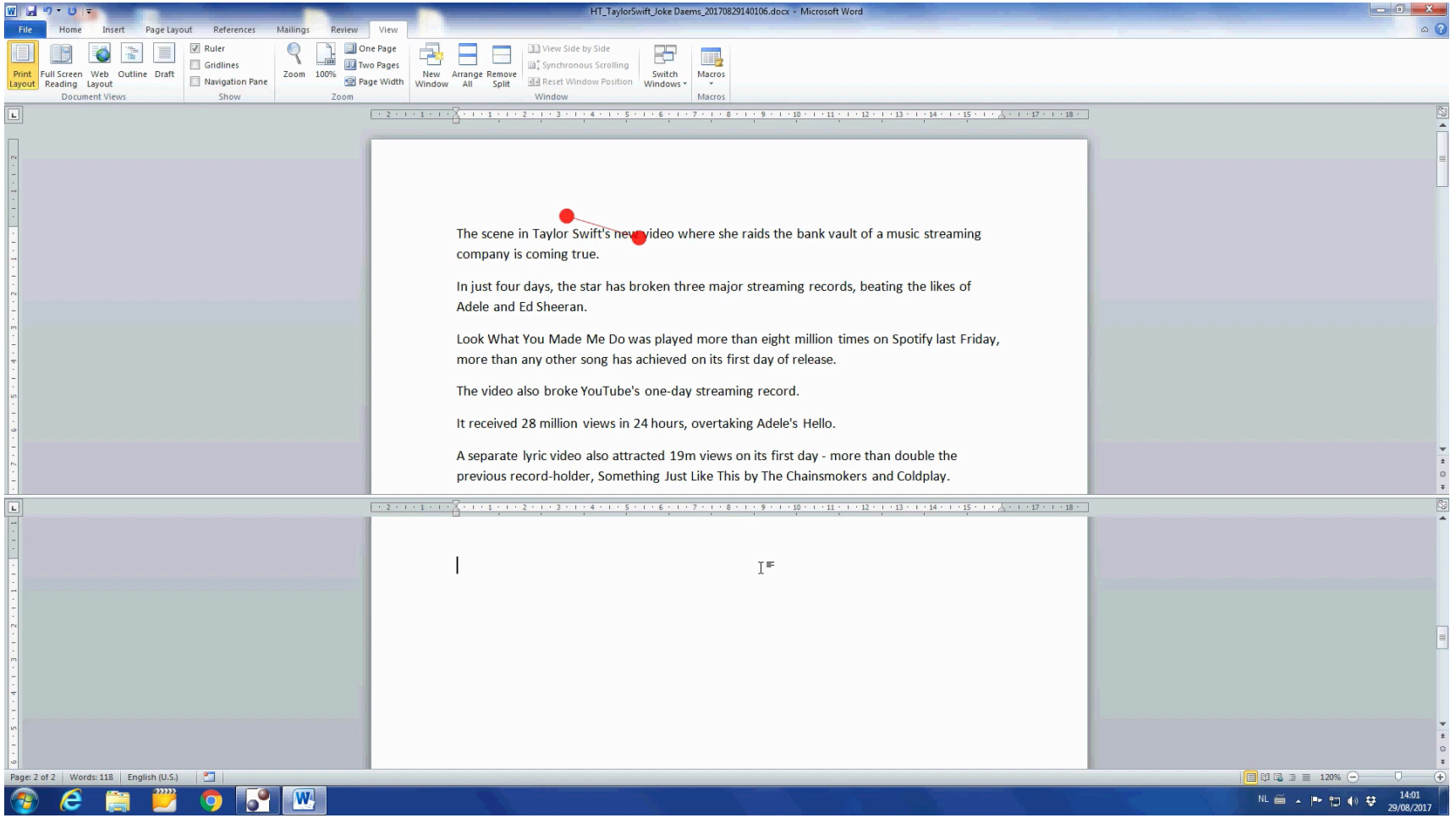
KEYSTROKE LOGGING

- Translation speed
- Pauses & pause patterns
- Insertions, deletions, revisions
- Production units (sequences of coherent typing activity)

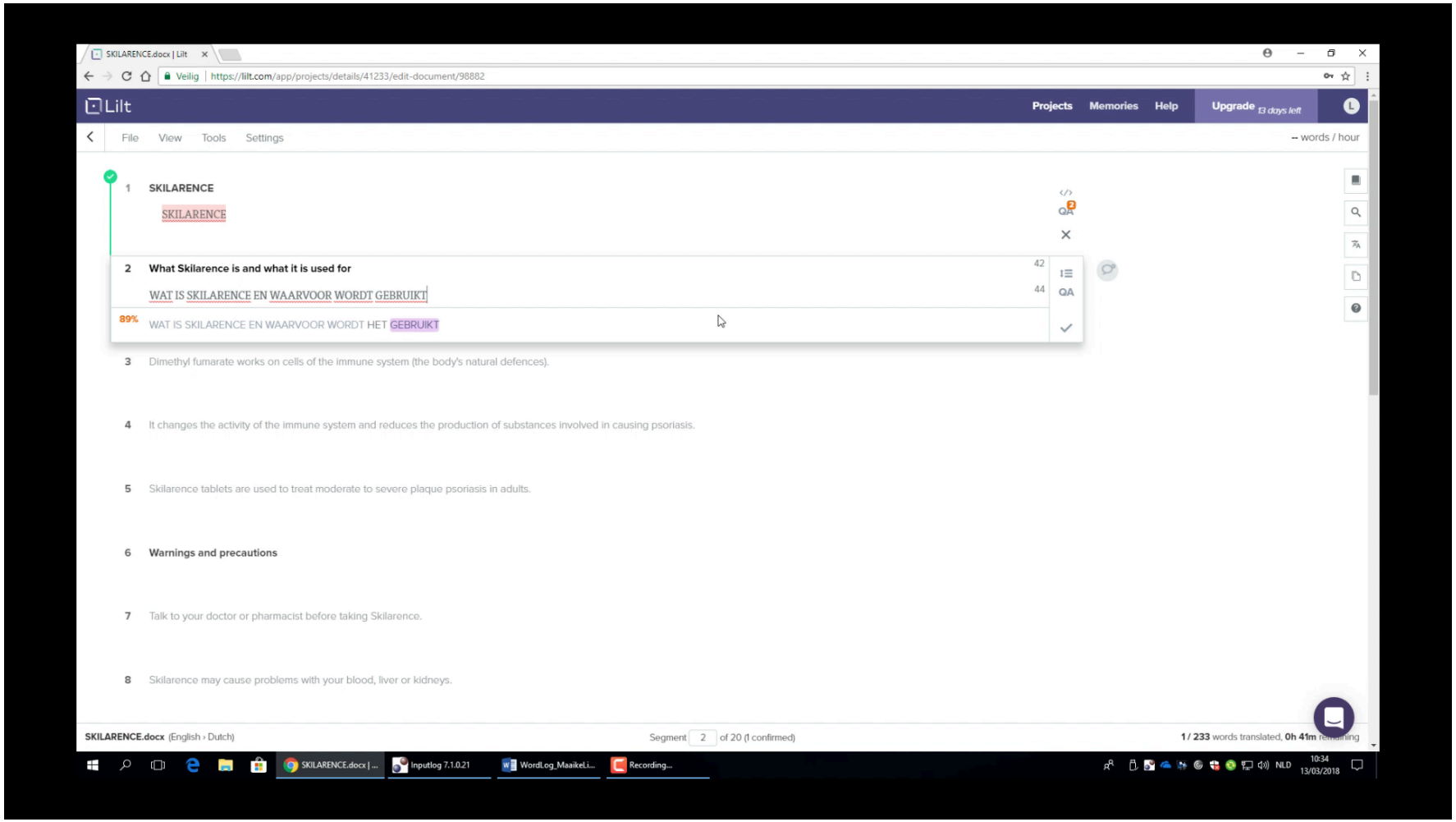
EYE-TRACKING

- Fixation time & pupil size → cognitive load
  - The longer the fixation and/or the larger the pupil, the more difficult the task.
- Fixations on source vs. target
- Regressions

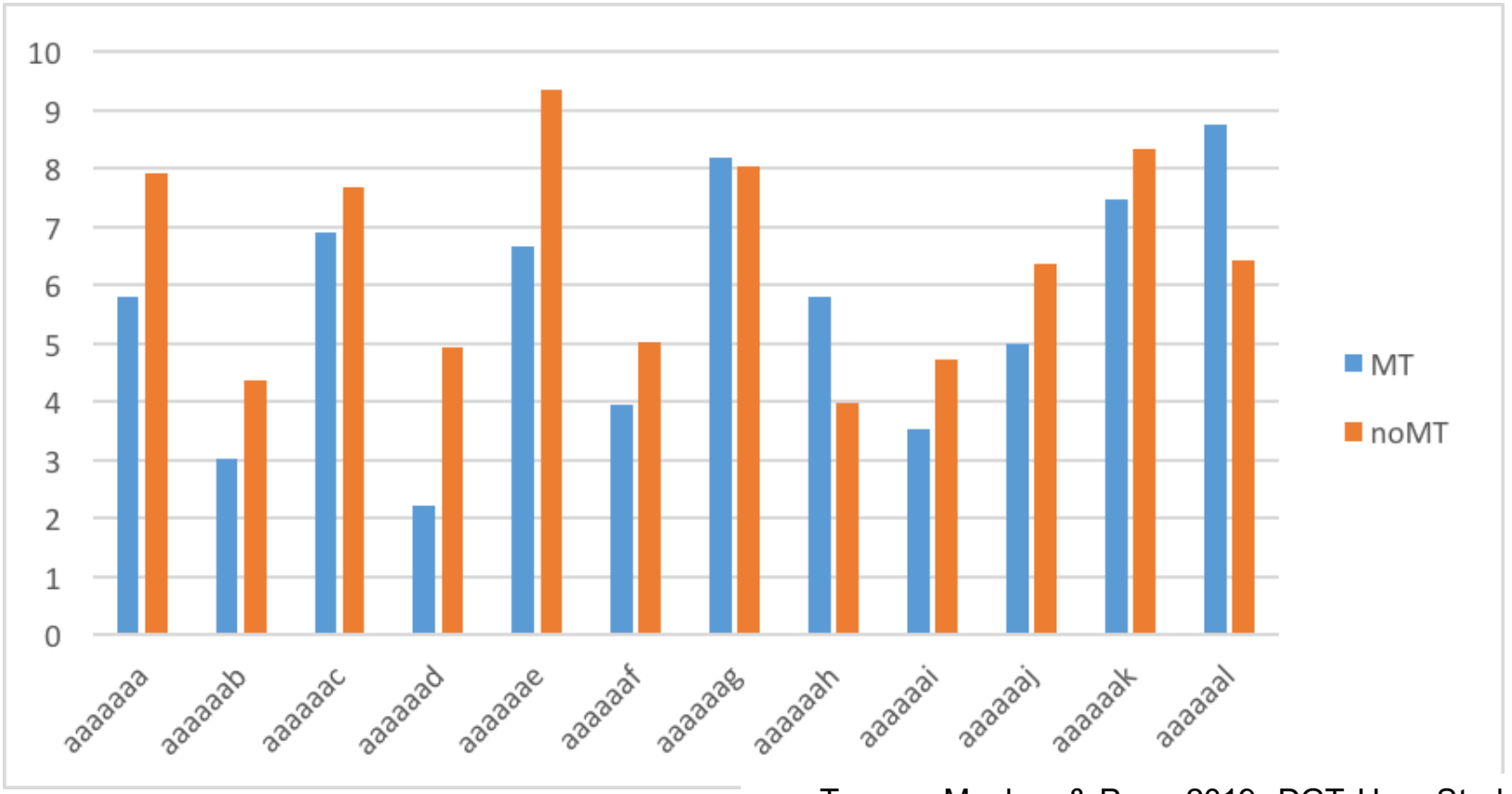
EYE-TRACKING



SCREEN CAPTURE

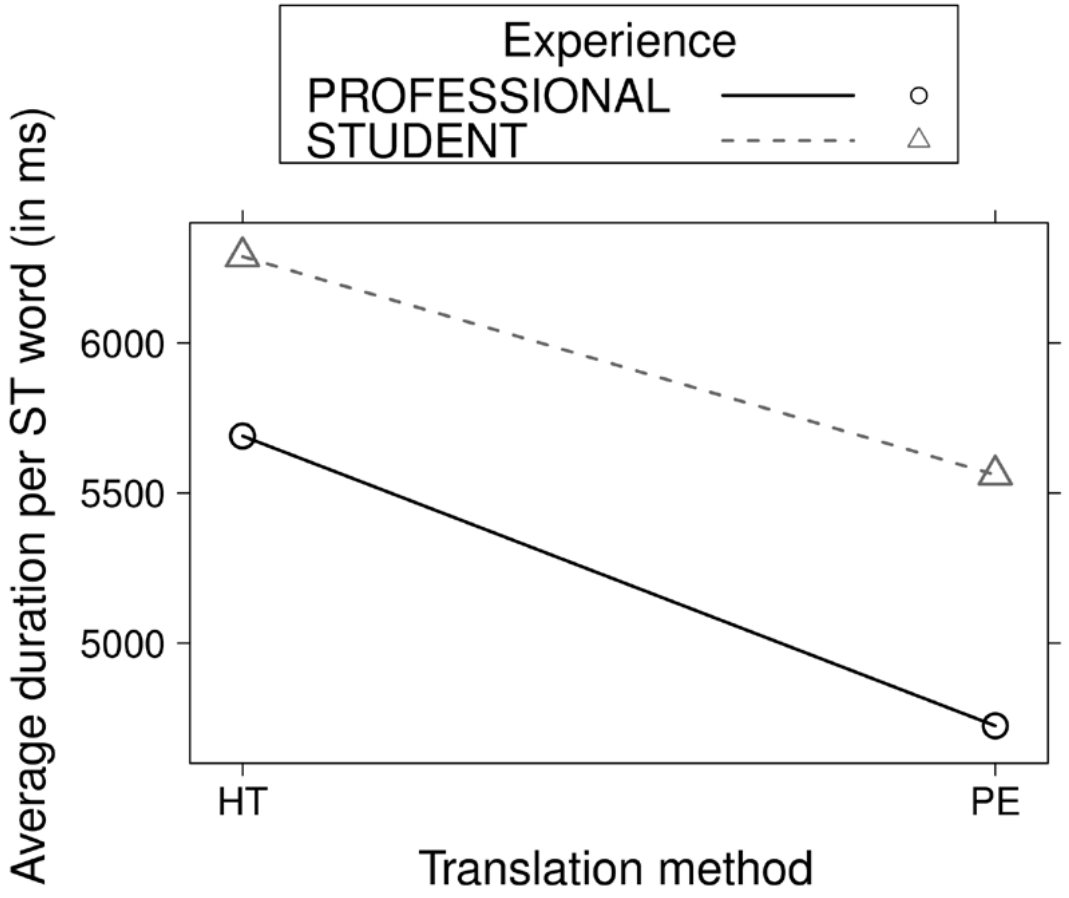


TRANSLATION SPEED: HT VS PE (DGT)



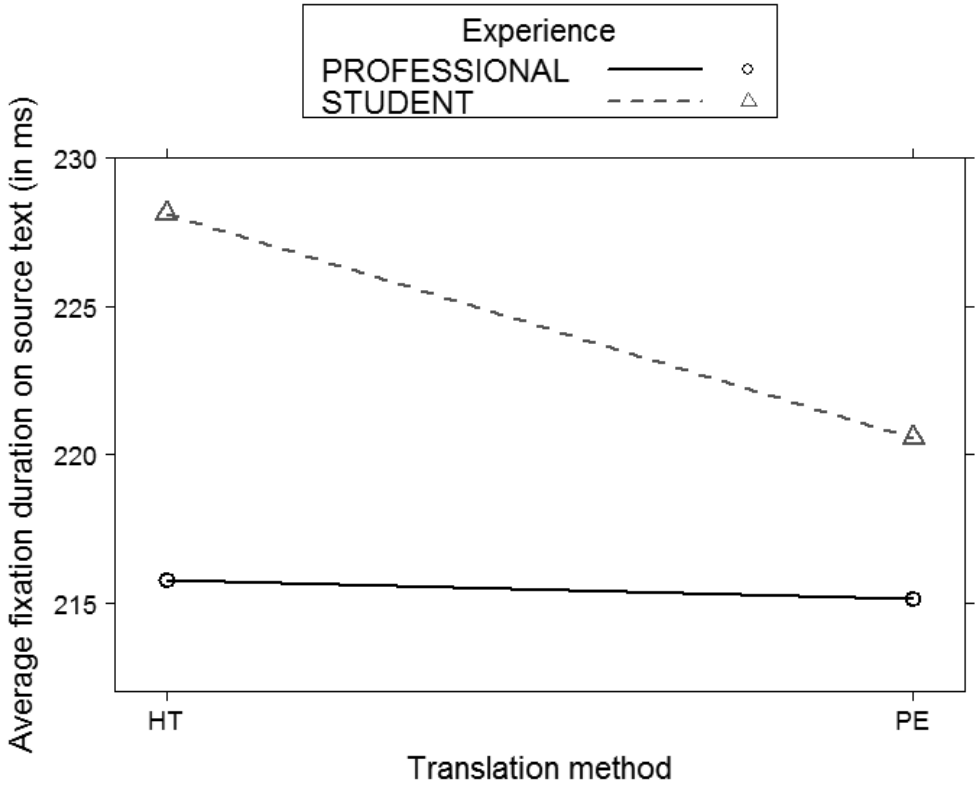
Tezcan, Macken & Prou. 2019. DGT User Study

TRANSLATION SPEED: HT VS PE (SMT)



Daems, Vandepitte, Hartsuiker & Macken. 2017. “Translation Methods and Experience ”. Meta

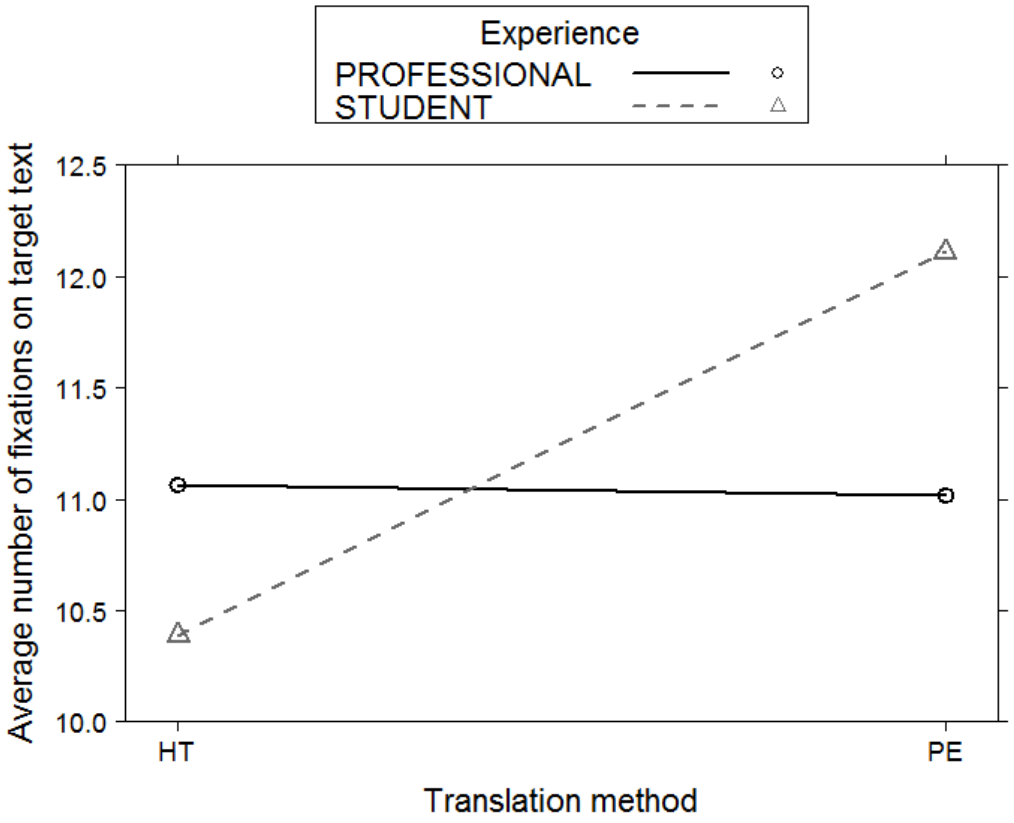
FIXATION DURATION SOURCE: HT VS PE



GHENT UNIVERSITY

Daems, Vandepitte, Hartsuiker & Macken. 2017. “Translation Methods and Experience”. Meta

NUMBER OF FIXATION TARGET: HT VS PE

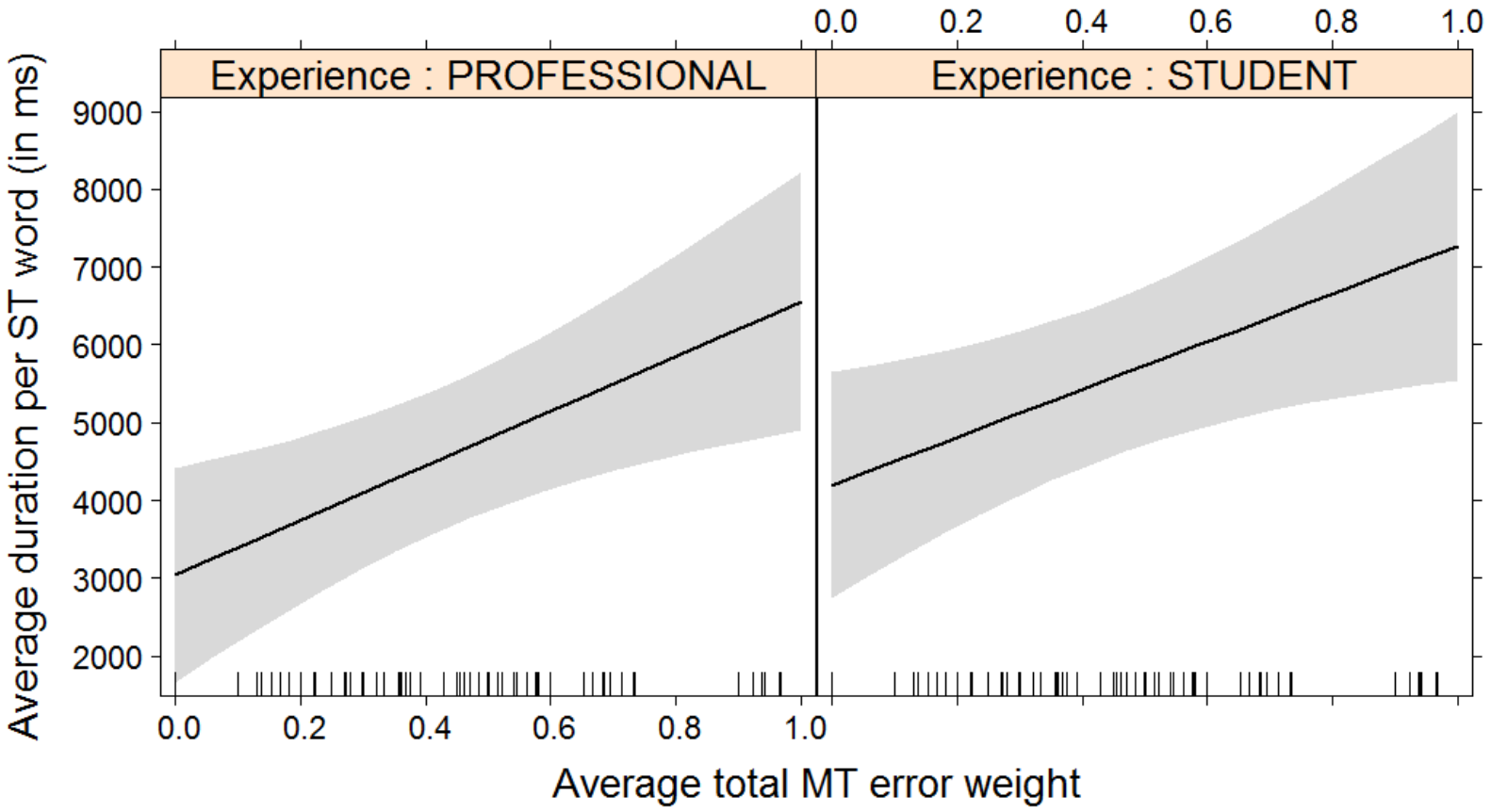


GHENT UNIVERSITY

Daems, Vandepitte, Hartsuiker & Macken. 2017. “Translation Methods and Experience”. Meta

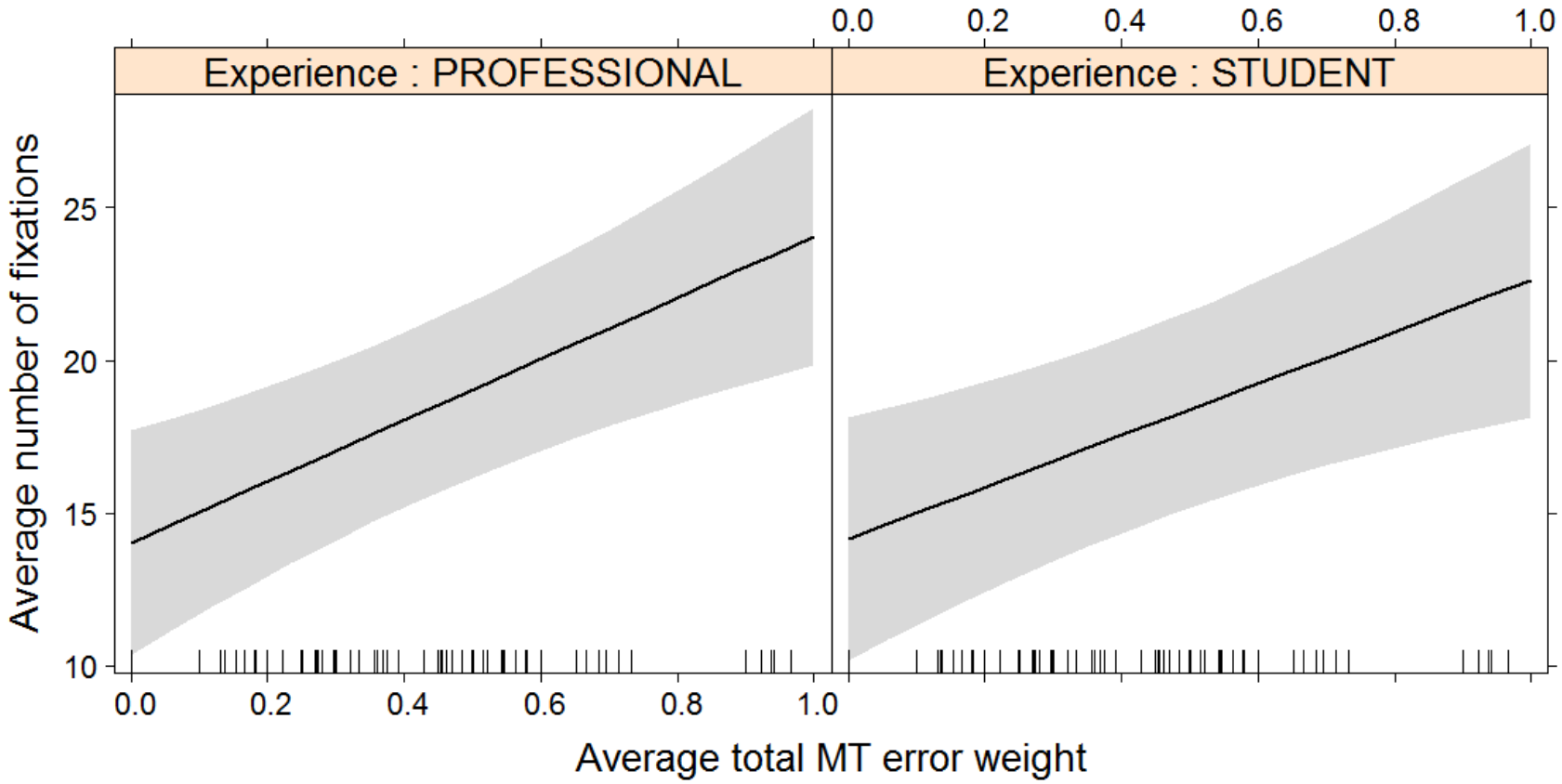
COMBINE PRODUCT & PROCESS DATA

AVERAGE MT ERROR WEIGHT ON DURATION



Daems, Vandepitte, Hartsuiker & Macken. 2017. “Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort.” Frontiers in Psychology.

AVERAGE MT ERROR WEIGHT ON FIXATIONS



Daems, Vandepitte, Hartsuiker & Macken. 2017. “Identifying the Machine Translation Error Types with the Greatest Impact on Post-editing Effort.” Frontiers in Psychology.

PREDICTING DIFFICULTY IN TRANSLATION

- Can we automatically predict whether a text is difficult to translate?
- Correlate product features with process features (proxy for cognitive effort)
  - Product = word translation entropy, syntactic equivalence
  - Process = pauses, revisions, fixations

HT: PRODUCT & PROCESS

Table 4. Correlations between word translation entropy (HTra) and process features.

	DURATION			REVISION				GAZE	
	AvgPauseRatio	Pausedur	Pdur	Mdel	Mins	Nedit	Scatter	FixS	FixT
prof	**−.1160**	.1854	**.1668**	**.1038**	**.2068**	.3729	(.0479)	**.1567**	.2011
stud	−.1119	**.1864**	.1338	.0930	.1576	**.4708**	(.0568)	.0991	(.0643)

Vanroy, De Clercq & Macken. 2019. “Correlating Process and Product Data to Get an Insight into Translation Difficulty.” Perspectives

HT: PRODUCT & PROCESS

Table 4. Correlations between word translation entropy (HTra) and process features.

	DURATION			REVISION				GAZE	
	AvgPauseRatio	Pausedur	Pdur	Mdel	Mins	Nedit	Scatter	FixS	FixT
prof	**−.1160**	.1854	**.1668**	**.1038**	**.2068**	.3729	(.0479)	**.1567**	.2011
stud	−.1119	**.1864**	.1338	.0930	.1576	**.4708**	(.0568)	.0991	(.0643)

Table 5. Correlations between syntactic equivalence (CrossS) and process features.

	DURATION			REVISION				GAZE	
	AvgPauseRatio	Pausedur	Pdur	Mdel	Mins	Nedit	Scatter	FixS	FixT
prof	**−.1526**	**.1482**	**.1901**	**.1371**	**.2661**	**.3098**	.0817	.1460	.2158
stud	−.1168	.1153	.0926	.0753	.1398	.1555	(−.0345)	(.0213)	(.0614)

Vanroy, De Clercq & Macken. 2019. “Correlating Process and Product Data to Get an Insight into Translation Difficulty.” Perspectives

PROJECTS

- DPC: Dutch Parallel Corpus
- ROBOT: A comparative study of process and quality of manual translation and the post-editing of machine translations
- SCATE: Smart Computer-Aided Translation Environment

- ArisToCAT: Assessing The Comprehensibility of Automatic Translations
- PreDicT: Predicting Difficulty in Translation
- Mutualist: Machine translation with User-specific Training and User-specific Adaptation for Literary texts

FACULTY OF ARTS AND PHILOSOPHY

DEPARTMENT OF TRANSLATION, INTERPRETING AND COMMUNICATION

HOW PRODUCT AND PROCESS DATA COMPLEMENT EACH OTHER IN TRANSLATION STUDIES

Lieve Macken, FLW Research Day, September 11th 2019

GHENT UNIVERSITY