TaLC2024

Capitalizing on the ChatGPT hype to rebrand the use of corpora in translator education: suggestions for a strategic shift from "corpus tools" to "data literacy"

RUDY LOOCK

Université de Lille

STL

STARTING POINT: CORPORA FOR TRANSLATION

- ✓ Monolingual / Bilingual (comparable + parallel)
- ✓ Online / DIY (Do-It-Yourself) with concordancer
- ✓ Improve source text understanding
- ✓ Improve target text writing
- ✓ Terminological extraction, glossary compiling

STARTING POINT: CORPORA FOR TRANSLATION

✓ And we know it works!

Bowker & Pearson (2002), Kübler (2003), Zanettin et al. (2003),Bowker (2011), Bernardini & Ferraresi (2013), Frankenberg-Garcia (2015), Loock (2016), Giampieri (2021), Kübler et al. (2022)…

⇒**Electronic corpora as translation, but also revision and post-editing tools**

15 Use the most relevant IT applications, including the full range of office software, and adapt rapidly to new tools and IT resources having critically assessed their relevance and the impact of change on their work practices

16 Make effective use of search engines, corpus-based tools, text analysis tools, computer- assisted translation (CAT) and quality assurance (QA) tools where appropriate

18 Understand the basics of MT systems and their impact on the translation process, and integrate MT into a translation workflow where appropriate

19 Recognise the importance and value of translation and language data, demonstrating data literacy

1 Analyse a source document, identify potential textual and cognitive difficulties and assess the strategies and resources needed to reformulate it in line with communicative needs

3 Carry out research to evaluate the relevance and reliability of information sources with regard to translation needs

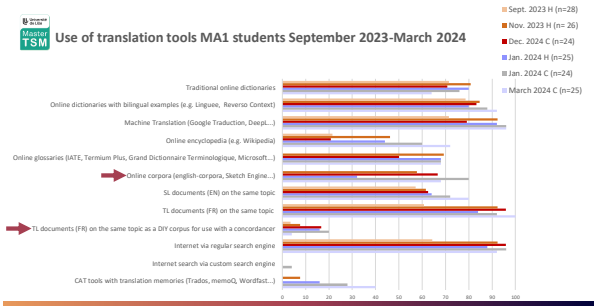
Most translation training programmes include corpus training
(Frérot & Karagouch 2016, Mikhailov 2022)

A LOVE STORY?

I ❤ corpora!

Think twice!

- ✓ Corpora = tools that students find difficult to master
- ✓ Gap with the professional world



Looking back on 10 years of corpus teaching in translator education:

MA1: corpora as translation aids

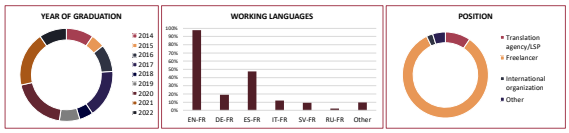
- online/personal [DIY] corpora for general/specialized data
- terminological/phraseological/collocational information
- monolingual/multilingual glossary compilation

MA2: corpora as tools for research

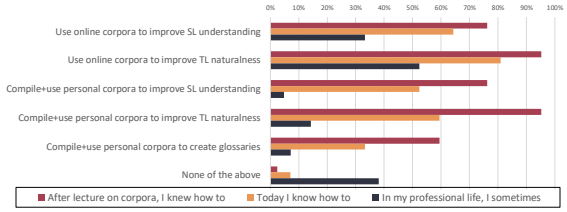
- analysis of translated texts by professionals/students; humans/machines
- link with translation quality

SURVEY FORMER STUDENTS

Online survey of students who graduated between 2014 and 2022
Feedback on their skills and use or corpus-based tools
42 respondents (March-April 2023) [response rate=27.6%]



SURVEY FORMER STUDENTS



SURVEY FORMER STUDENTS

"I use Sketch engine to find matches in the target language to make my translations more fluent."

"I use corpora with AntConc mainly to get familiar with the terminology of a particular client/topic by compiling relevant articles or files"

"I mainly use corpora for very large or very technical projects."

"I often compile mini-corpora in the target language for technical translations and I often use online corpora in source and target languages."

"I use corpora on a daily basis, those available online."

"In my opinion, the use of corpora makes all the difference in a translation. It really helps to go from an acceptable quality to a fine and careful work based on informed choices (and this is not an answer to please you!)"

SURVEY FORMER STUDENTS

*"We **don't have time** to use concordancers, because we don't have time to search for texts for both languages. I never used this since 2014, though it looked great..."*

*"The **deadlines** requested by clients almost never allow me to take the time to use corpora. In addition, the use of quality online corpora is **not free**."*

*"I don't use corpora in my daily tasks because the **expected pace** does not allow for such extensive research."*

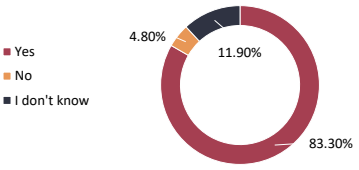
*"I use them for each project. Very often bilingual corpora (...) but I don't compile corpora with AntConc or Bootcat. I like SketchEngine very much, but the **price** is too high."*

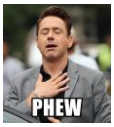
*"Generally, clients provide **TMs that are sufficient** and I don't need additional corpora. Occasionally, for rare terms or expressions, I search online corpora."*

*"Compiling a corpus offline and running it with AntConc is too **time consuming**. Online corpora provide answers (or not) very quickly"*

SURVEY FORMER STUDENTS

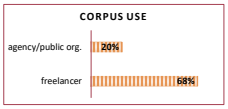
I am convinced that corpora can be used as translation aids



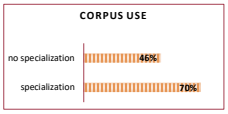


SURVEY FORMER STUDENTS

- Link with freelance vs. translation agency/public organization?



- Link with specialization? (law, medicine, IT, new technologies, marketing, tourism…)



ON THE MARKET

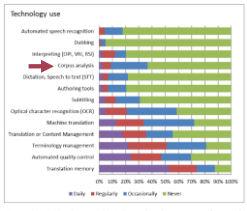


Figure 42 - Technology use by independent language professionals

Source: ELIS report 2022

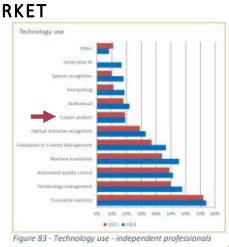
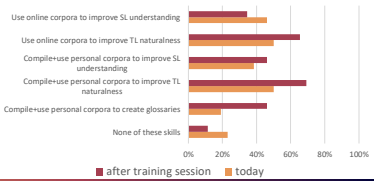


Figure 83 - Technology use - independent professionals

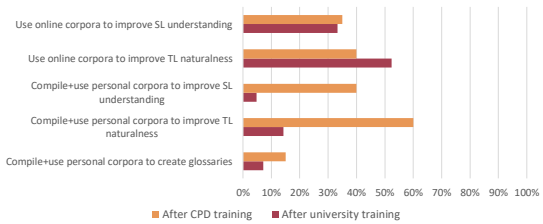
ELIS 2024

ON THE MARKET

Continuous Professional Development (5 sessions 2019-2023)



CORPUS USES AFTER UNIVERSITY/CPD TRAINING



PROPOSALS

#1 Focus teaching on **data literacy** and a general approach of translation tools

- The **role of data** in **all the tools used by translators** (online dictionaries, CAT tools, MT, GenAI tools…)
- **Data literacy** = "ability to collect, manage, evaluate, and apply data, in a critical manner" (Ridsdale et al. 2015: 11)
- Instead of presenting tools separately, **unify** them as **linguistic databases searchable via different types of interfaces** (web/software)

https://mashable.com/article/google-search-low-quality-research

PROPOSALS

#1 Focus teaching on **data literacy** and a general approach of translation tools

⇒ The **ChatGPT hype** provides a fantastic opportunity

- Everyone has heard of it (1m users in 5 days)
- Lots of discussion on data (quality, origin, accuracy)
- Ethical debates (intellectual property)
- Words like "hallucinations" are now mainstream

PROPOSALS

#2 Replace the word corpus with **linguistic data(base)** *(please don't hit me!)*

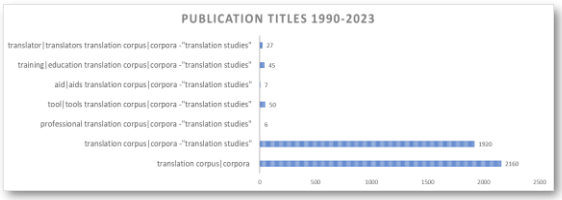
- **Terminological gap** with industry (corpus = academic research)
- Term **absent** from job/internship adverts (vs. data)
- Some CAT tools features < corpus methodology (RWS Trados/memoQ's LiveDocs), but the word *corpus* almost never used
- **Not without problems**: some online tools/software use the word corpus (english-corpora website, Sketch Engine, concordancers)

PROPOSALS

#3 Separate corpora as **translation tools** from research tools in translation studies

- **Bias** in translation studies literature and also corpus teaching
  - Most translation training programmes introduce corpora with a research orientation (Mikhailov 2022)
  - Google Scholar shows *translation* and *corpus/corpora* appear in more than 2,000 titles of publication, among which very few include *translator(s), aid(s)*, or *tool(s)*

PROPOSALS



Only Google Scholar, titles, English language (Jan. 2023)

PROPOSALS

#4 Set up a **metacognitive approach** with students

- metacognition (Flavell 1979) = awareness or analysis of one's own learning or thinking processes
- at the end of each MA2 class, students write 5-10 lines to make them bridge the gap between training and professional life:

What can you pick up from today's class that you can apply in your future professional life?

I know how... | We have seen... | I have learnt... → I will be able/need to... | I will pay attention to...

Analysis of a **metacognition exercise corpus**:
- September-November 2023 (7 lectures)
- 22 students enrolled, 18 attending on average
- 142 feedbacks | 20,444 tokens (144 on average)

Analysis of a **metacognition exercise corpus**:
- September-November 2023 (7 lectures)
- 22 students enrolled, 18 attending on average
- 142 feedbacks | 20,444 tokens (144 on average)

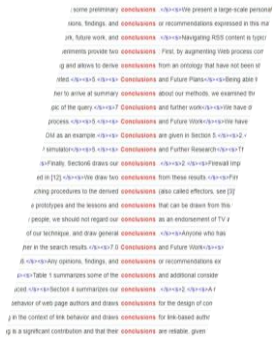
2-grams

3-grams

PROPOSALS

#5 Use simple(r) tools
- Many tools today are simple to use: online translators, dictionaries, search engines… but corpus tools not always
- See Crosthwaite & Baisa (2004) for DDL: most corpus tools are complex, research-oriented, intimidating for teachers and learners.
- Why not 2 versions of the same tool? *(again, don't hit me!)*
- More integration in CAT tools?

CONCLUSIONS



Since there's a lot of talk about data these days, let's talk about data!

- By **unifying** the approach to translation tools, most of which need data
- By focusing on **data education/literacy**
- By focusing more on **professional use and jargon**
- By keeping things as **simple** as possible

Contact: rudy.loock@univ-lille.fr @RudyLoock