

From corpus linguistics to AI: The enduring importance of having the right data

Lynne Bowker (Lynne.Bowker.1@ulaval.ca)

XVI Encontro de Linguística de Corpus (ELC)

University of Brasilia, 22-24 October 2024

Canada Research Chair
in Translation, Technologies,
and Society



UNIVERSITÉ
LAVAL

Chaire de recherche du Canada
en traduction, technologies
et société



UNIVERSITÉ
LAVAL

Corpus design (1990s)

- Very active research area!
 - Not a random assortment of texts
 - Planned and motivated rather than opportunistic
 - Copyright issues
 - Sampling, balance, representativeness

Images: academic.oup.com

JOURNAL ARTICLE

Corpus Design Criteria [Get access ↗](#)

SUE ATKINS, JEREMY CLEAR ✉, NICHOLAS OSTLER

Literary and Linguistic Computing, Volume 7, Issue 1, 1992, Pages 1–16,
<https://doi.org/10.1093/llc/7.1.1>

Published: 01 January 1992

JOURNAL ARTICLE

Representativeness in Corpus Design

[Get access >](#)

DOUGLAS BIBER ✉

Literary and Linguistic Computing, Volume 8, Issue 4, 1993, Pages
243–257, <https://doi.org/10.1093/llc/8.4.243>

Published: 01 January 1993

JOURNAL ARTICLE

Longman/Lancaster English Language Corpus – Criteria and Design [Get access >](#)

Della Summers

International Journal of Lexicography, Volume 6, Issue 3, Autumn
1993, Pages 181–208, <https://doi.org/10.1093/ijl/6.3.181>

Published: 01 October 1993

PDF

Key features emerged

“...a **large** collection of **authentic** texts that have been gathered in **electronic form** according to a **specific set of criteria**.” (*Bowker and Pearson 2002*)

- Machine-readable form
- Size
- Specific criteria
- Authenticity

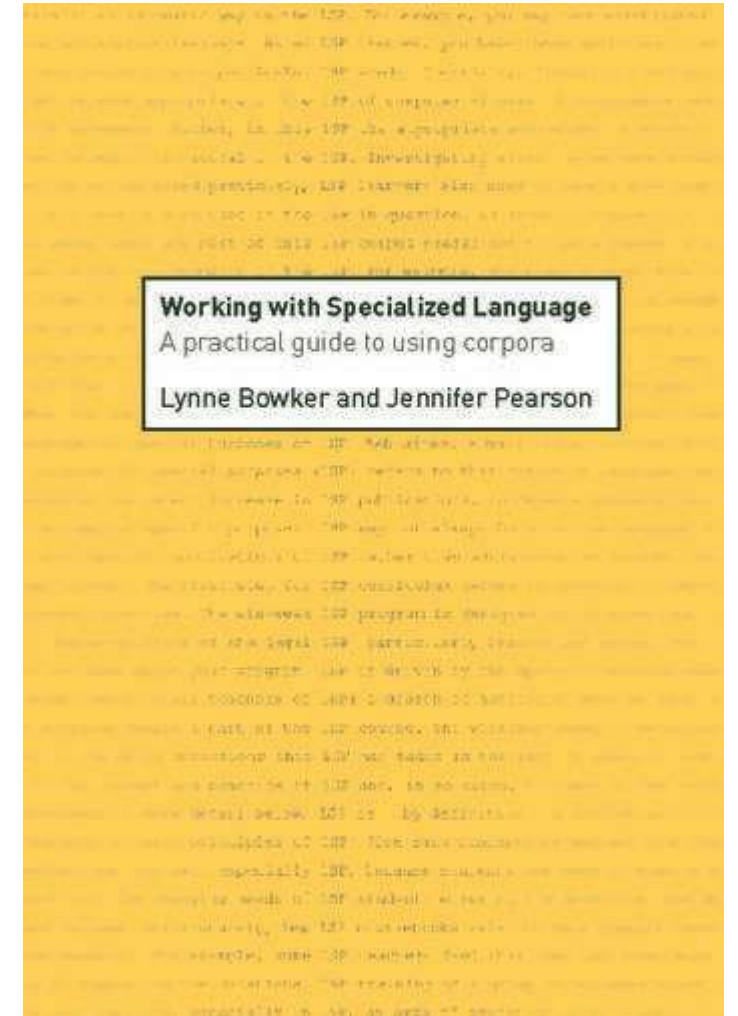


Image: rouledge.com/

Many different types of corpora

Designing Learner Corpora

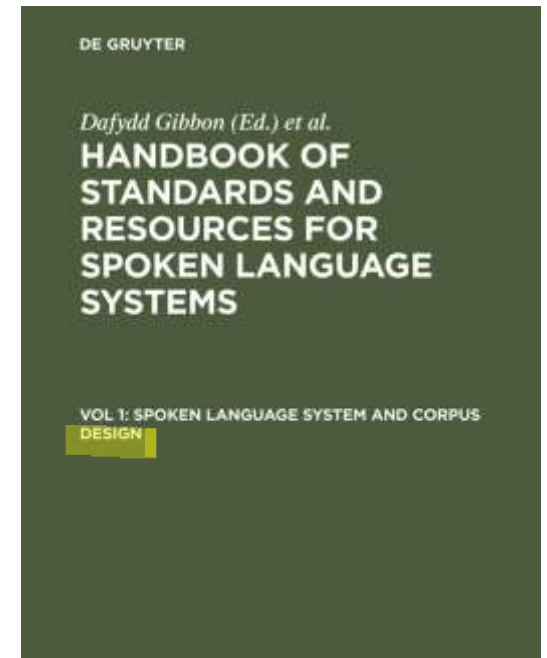
Collection, Transcription, and Annotation

By *Philippa Bell, Caroline Payant*

Book [The Routledge Handbook of Second Language Acquisition and Corpora](#)

Edition 1st Edition

First Published 2020



JOURNAL ARTICLE

Designing a Corpus for Translation and Language Teaching: The CEXI Experience

Silvia Bernardini

TESOL Quarterly

Vol. 37, No. 3 (Autumn, 2003), pp. 528-537 (10 pages)

Published By: Teachers of English to Speakers of Other Languages, Inc. (TESOL)





Procedia - Social and Behavioral Sciences

Volume 173, 13 February 2015, Pages 293-299



Corpus Design and Compilation Process for the Preparation of a Bilingual Glossary (English-Spanish) in the Logistics and Maritime Transport Field: LogisTRANS ☆

María Araceli Losey-León  

Still relevant today! (2023)

Sustainability & Health Corpus

The Sustainability & Health Corpus is housed at the Centre for Healthcare Education (SHE), Faculty of Medicine, University of Oslo. The SHE Corpus is a large suite of electronic corpora of health and health-related texts, accompanied by an open access software interface and visualization tools to support research in the fields of healthcare and medical humanities.

Team including Mona Baker, Gabriela Saldanha, Henry Jones, Jan Buts, and others

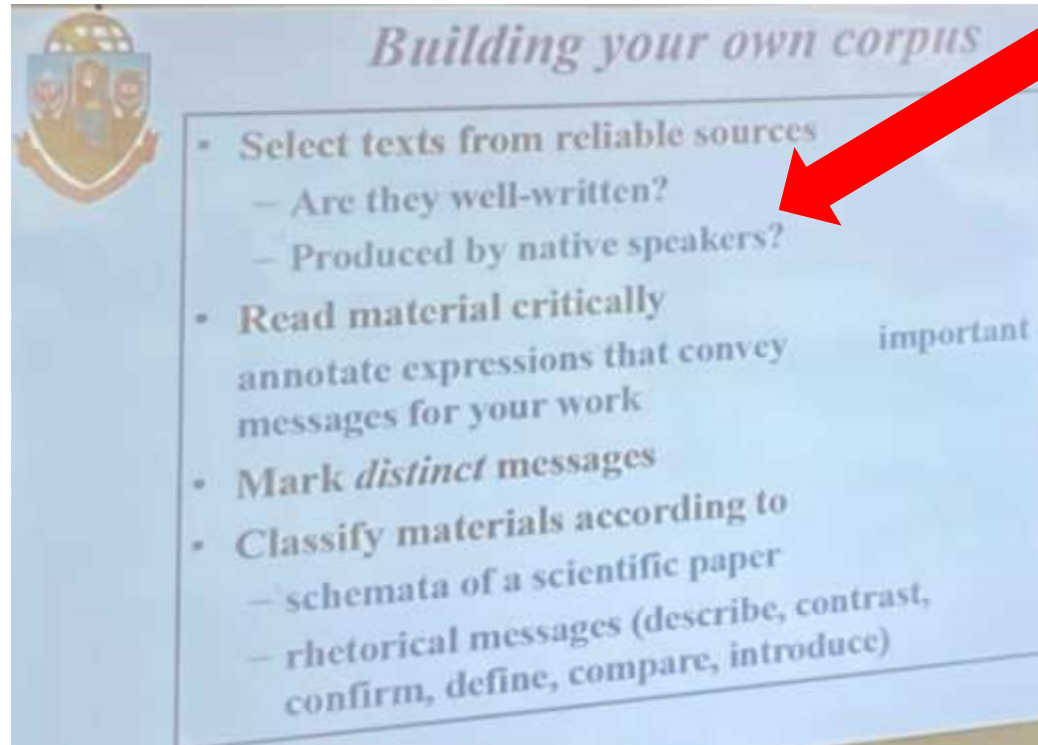
Images: <https://www.shecorpus.net/corpus-design-selection-criteria/>

Corpus Design & Selection Criteria

This document outlines the main principles adopted to ensure the integrity and continued relevance of the Sustainability & Health Corpus to a wide range of researchers in the broad field of medicine and healthcare, and to those interested in examining the intersection of health and sustainability.

- [Corpus Design](#)
- [Representativeness and Balance](#)
- [Selection Criteria](#)
 - [External](#)
 - [Source](#)
 - [Document Format](#)
 - [Time Span](#)
 - [Region](#)
 - [Copyright Status](#)
 - [Internal](#)
 - [Topic area](#)
 - [Region](#)
 - [Text and graphics](#)
- [Guiding Principles for Purposeful Sampling](#)

Even in October 2024!!

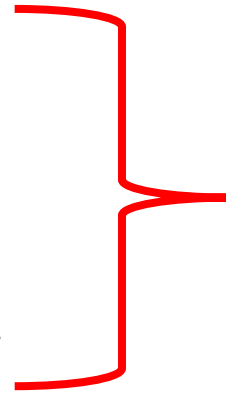


- Workshop by Osvaldo Novais de Oliveira Jr (21 October 2024)

From standalone to integrated resources

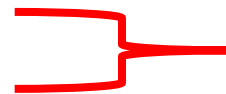
- Corpora are processed by

- Concordancers
- Term extractors
- Translation memories
- Machine translation systems
 - Corpus-based MT: EBMT, SMT, NMT



General features of a corpus remained relatively stable

- Generative AI tools



Some shifts in corpus features

Machine-readable form

- More important than ever!
- AI tools process corpora more intensively than other tools
 - e.g. concordancers and translation memories mainly do superficial pattern matching
 - AI tools go further, *attempting* to analyze information and present fully formed solutions (e.g. texts, translations)



...y have a different **pattern** than traffic killings of
exhibit a different **pattern** .
on and landscape **pattern** on American martens
the tree diversity **pattern** in Europe be generate
)) winter feeding **pattern** and Zostera sources a
Marine pollution **pattern** of Skagerrak and Katt

Image: alf.hum.ku.dk/korp/

Size

- Notion of “large” has evolved
 - Brown corpus = 1 million words
 - British National Corpus = 100 million words
 - ChatGPT-3.5 = 300 billion words (Hughes 2023)
 - ~90% in American Eng., ~10% in all other languages, pivot translation
- We’ve moved *way* beyond “Big Data”
 - Enormous, colossal, gargantuan!
- How is this need for **extremely** large corpora affecting other corpus features?



Image: pixabay.com/

Size influences performance & output quality

- Availability of (machine-readable) texts differs from one language to the next
- Disparities in tool performance between high- and low-resource
 - Languages
 - Language varieties
 - Domains



Image: pixabay.com/

Specific criteria

- The best corpora contain texts that have been selected according to **relevant** criteria
 - e.g. in translation, one important criteria is translation **quality**
- Implies being **selective**, but selectivity may not be compatible with the need for **colossal** data
 - Opportunistic, sometimes unethical
 - Garbage in, garbage out...



Image: pixabay.com/

Where do AI training data come from?

- Most of the searchable internet (including copyrighted materials)
 - i.e., anything not behind a login page
 - Criteria = **accessible**...
- Crawlers
 - Travel from link to link and index location of info in database
- Scrapers
 - Download and extract the info
- Public profiles, blogs, personal webpages, company sites, online marketplaces, voter registration databases, government webpages, Wikipedia, Reddit, research repositories, news outlets, academic institutions, etc.
- Items may be individually worthy, but collectively not coherent (alphabet soup!)

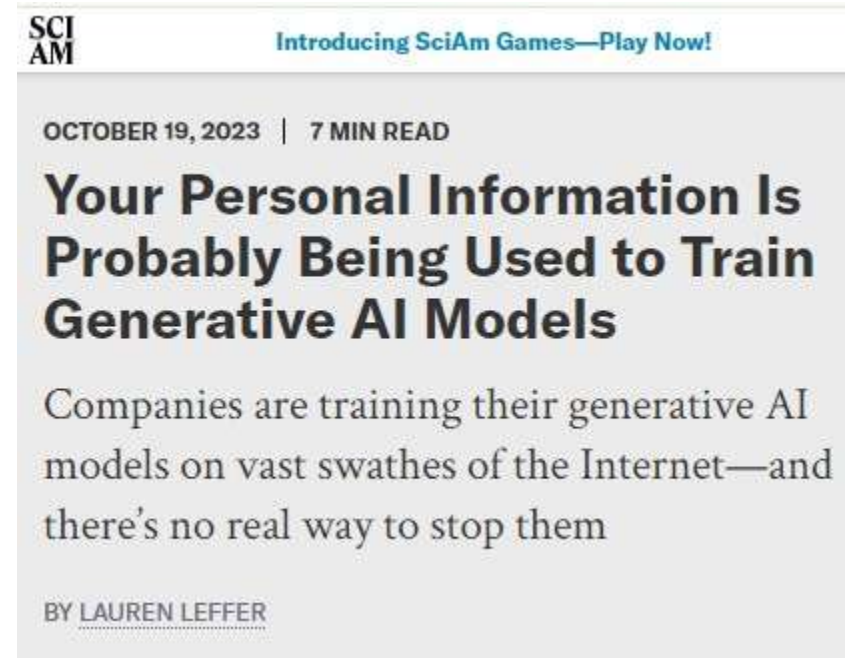


Image: <https://www.scientificamerican.com/>

Authenticity

- Formerly **sacrosanct**!
- McEnery (2022)
 - “a large body of linguistic evidence composed of **attested** language use” (494)
 - “a collection of **naturally occurring** language data” (495)
- In the age of colossal data, authentic high-quality texts are **valuable** commodities
 - High-quality usu. means human-created (or even professionally created)



Image: pixabay.com/

Good quality

- As observed by Kenny (2011: 2), the reason that the developers of translation tools use corpora of human translations to train their systems is because
“such corpora are assumed to contain good answers to translation problems; and they are assumed to contain good answers **precisely because they contain translations performed by human beings.**”



Image: pixabay.com/

New York Times sues OpenAI, Microsoft for copyright infringement

Several generative AI companies facing similar suits by writers, visual artists

Thomson Reuters ·

Posted: Dec 27, 2023 4:58 PM EST | Last Updated: December 27, 2023



OpenAI, along with Microsoft, is being sued by the New York Times, which alleges the artificial intelligence company has cost it 'billions of dollars' in damages by illegally copying and using its works. (Dado Ruvic/Illustration/Reuters)

Image: <https://www.cbc.ca>

Academic authors 'shocked' after Taylor & Francis sells access to their research to Microsoft AI

NEWS JUL 19, 2024 BY MATILDA BATTERSBY

Image: <https://www.thebookseller.com/>

The image is a screenshot of a blog post from "The Scholarly Kitchen". The header includes the site name "THE SCHOLARLY kitchen" and navigation links: "ABOUT", "ARCHIVES", "COLLECTIONS", "TRANSLATIONS", "CHEFS", "PODCAST", and a "FOLLOW" button. The main title of the post is "Tracking the Licensing of Scholarly Content to LLMs". Below the title, it says "By ROGER C. SCHONFELD | OCT 15, 2024 | 8 COMMENTS". There are social media sharing buttons for X, Facebook, and LinkedIn, along with a "PRINT THIS PAGE" link. The main text of the post begins: "In recent months, several publishers have announced that they are licensing their scholarly content for use as training data for LLMs (Large Language Models). These deals illuminate how major publishers are grappling with their strategy amid uncertainty, but thus far they have been unavailable to smaller and medium size publishers. To understand the dynamics around this fast-developing market, my colleagues Maya Dayan and Dylan Ruediger and I are launching a tracker of these licensing deals." On the right side, there is a section titled "OFFICIAL BLOG OF:" followed by the logo for "Society for Scholarly Publishing" and a section titled "THE CHEFS" with four small profile pictures.

Image: <https://scholarlykitchen.sspnet.org/>

But there's still not enough authentic data!

- **Augmented** and **synthetic** data
- Ouroboros effect
 - AI models trained on AI-generated data
 - Amplification of bias, cycle of mediocrity, model collapse
 - Drop in quality



VICE

<https://www.vice.com/article/a-shocking-amount-of-...>

A 'Shocking' Amount of the Web Is Already AI-Translated ...

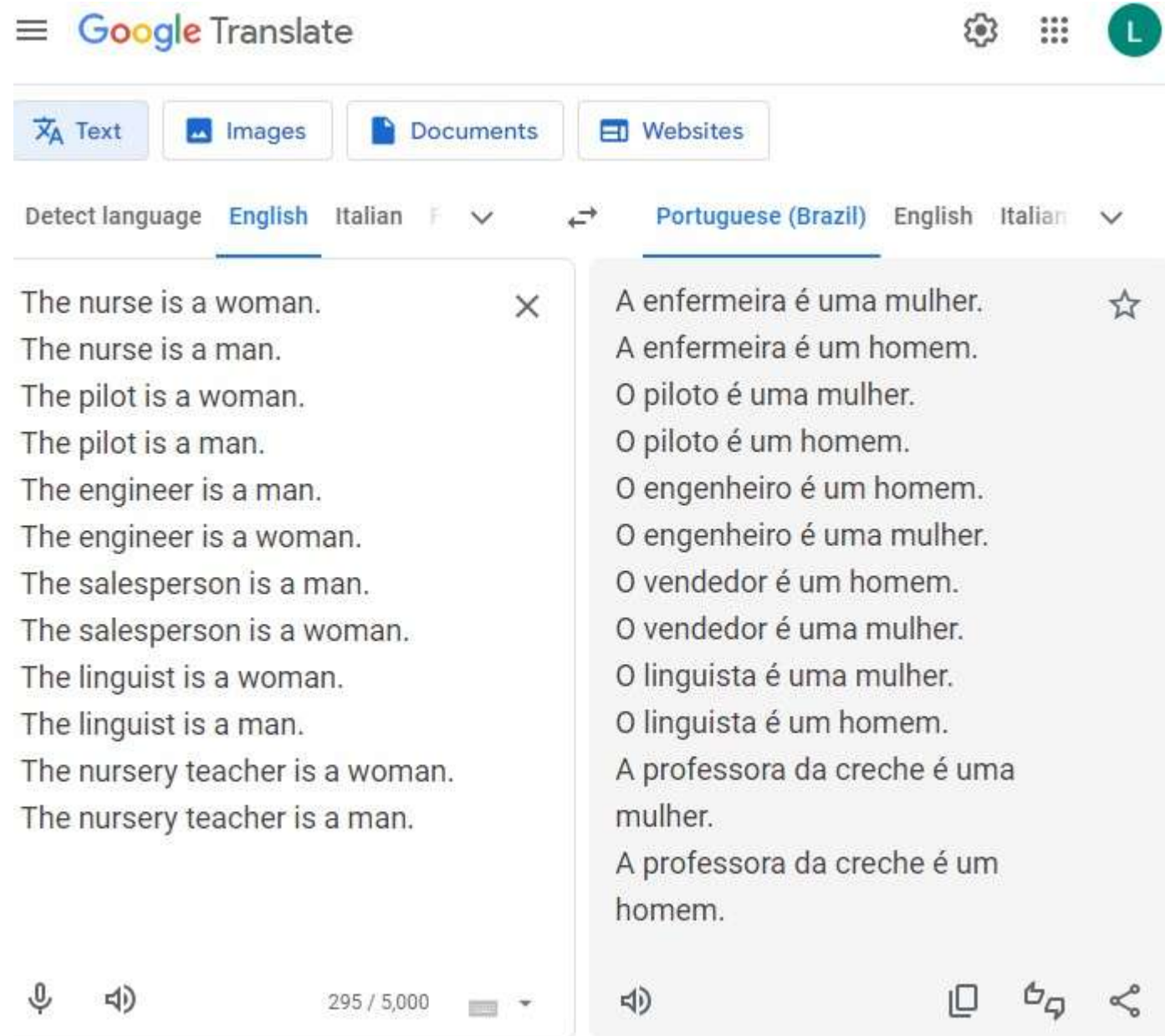
Jan 17, 2024 — A “shocking” amount of the internet is **machine-translated garbage**, particularly in languages spoken in Africa and the Global South, a new study has found.



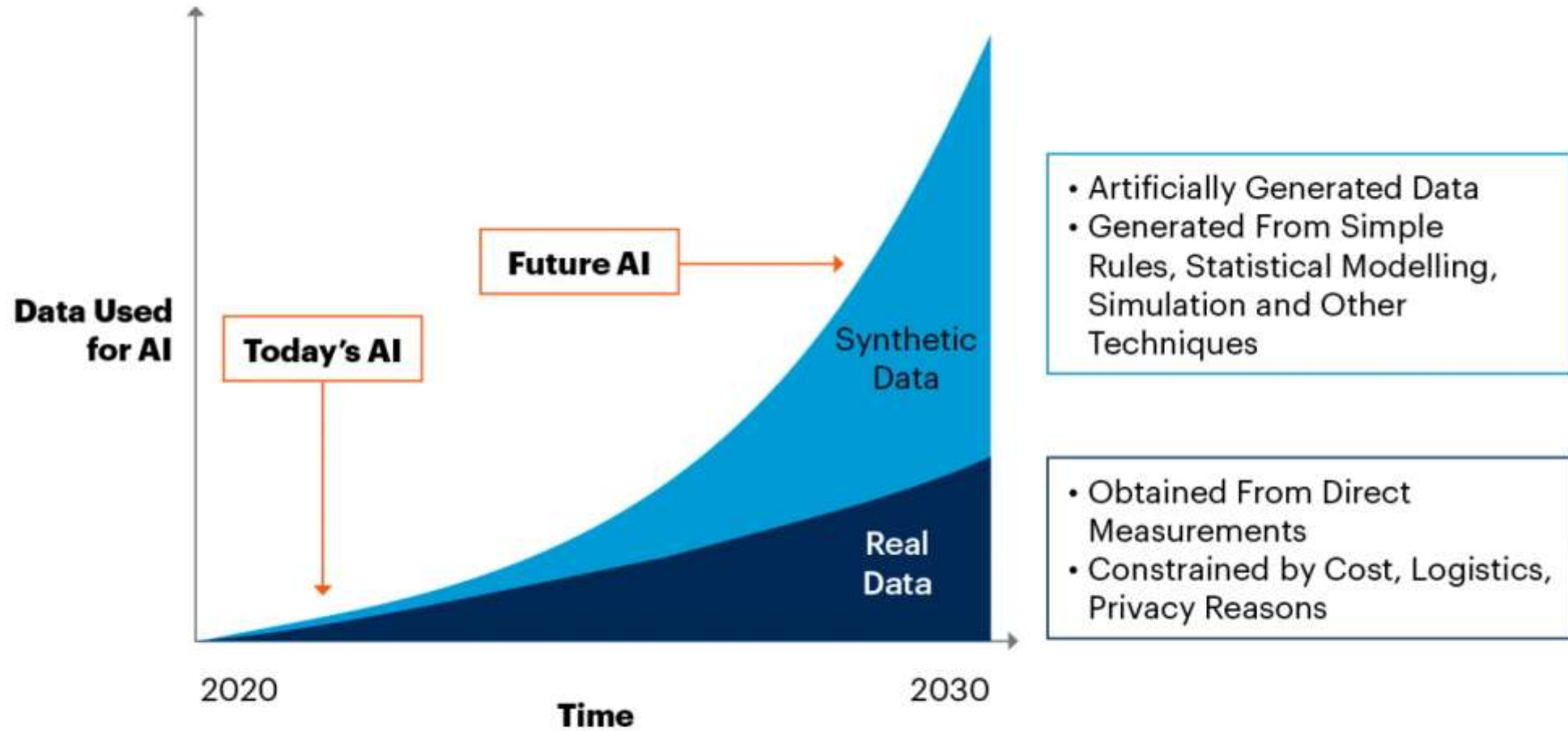
Image: pixabay.com/

E.g. nonsense and amplification of gender stereotypes perpetuated by Google Translate

Image: <https://translate.google.com/>



By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models



Source: Gartner
750175_C

Image: <https://www.enago.com/academy/>

Quick recap

- Machine-readable
 - Still relevant
 - Easier to obtain for some languages than others
- Size
 - Large is an understatement
 - Sampling no longer possible
- Specific criteria
 - Now incompatible with size
 - Less selectivity, more opportunism, questionable practices
- Authenticity
 - Now incompatible with size
 - At risk... in danger of extinction!

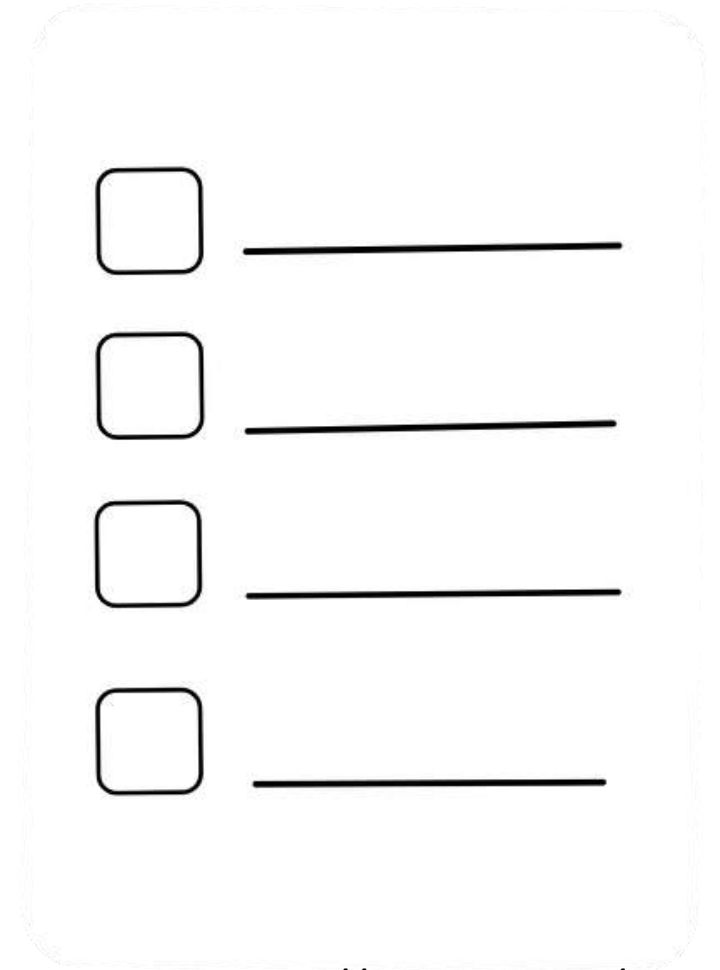


Image: <https://pixabay.com/>

Given these changes...

- Are corpora in the age of AI still corpora...?
- Is the notion of corpus becoming a “zombie concept” (Beck 2001)?
 - Sociology
 - Categories that no longer apply to the world but are continuously perpetuated (kept artificially alive)
 - Nation state, social class, family
- “Zombie concepts” can prevent us from properly developing new concepts or terminology



Image: <https://pixabay.com/>

R. Loock (TALC 2004): “rebrand” corpora in translator training

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)

A terminological & conceptual conundrum!

- Do we want to **rebrand**?
 - ~~Corpus~~ → ???
- Do we want to **maintain** the term and encourage a return to the features that we originally associated with corpora?
 - Corpus → [“zombie concept”] → corpus
- Do we need to **split** into two concepts?
 - Corpus *and* ???



Image: <https://pixabay.com/>

Is bigger always better? Or necessary?

Large Language Models (LLMs)

- Multipurpose
- Need more training data & time, storage, computational resources
- Harder to customize and fine tune
- Only large corporations can build and operate them (social, eco cost)

Small is the new big: The rise of small language models



Sunita Tiwary
Jul 22, 2024

Small(er) Language Models (SLMs)

- Task specific
- Need less training data & time, less storage, fewer computational resources (more efficient + eco-friendly)
- Easier to customize and fine tune
- Smaller organizations can build and operate them (and control privacy more easily)

Image: <https://www.capgemini.com/>

SMLs require **curation**

- Quantity cannot always compensate for poor quality
- A return to the principles of good corpus design (best practices)
 - Selectivity
 - Quality (authenticity)
 - Relevance (specific criteria)
 - Size (less colossal)
 - Machine-readable (still need it ;-)
- Precedents
 - Translation memories (organized by client)
 - Domain-specific MT
 - Corpora for language varieties
 - ...

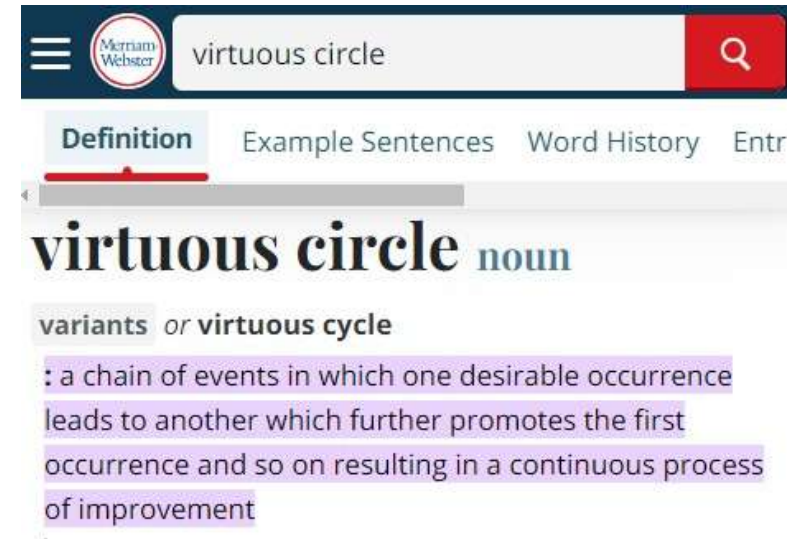


Image: www.merriam-webster.com

Data curator: a job for the future (or NOW!)

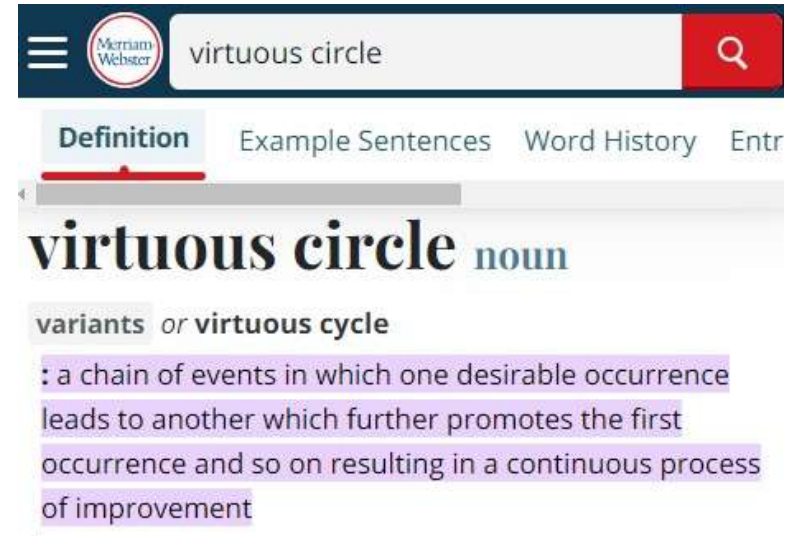
- Who would be good at this job?
- Someone who understands:
 - Text typology
 - Target audiences (KYA)
 - Localization (language varieties)
 - Balance
 - Representativeness
 - Pre-processing (e.g. cleaning, labelling)
 - Ethics
- Tech-savvy (corpus) linguists are more suited to this work than engineers or computer scientists



Image: <https://www.redbubble.com/>

Summing up

- Corpora in the age of AI no longer resemble the corpora used by corpus linguists
 - Size, design criteria, authenticity look different today
 - At what point should we stop calling them corpora?
- LLMs may be unsustainable...
- SLMs offer a number of advantages, but require careful curation and a return to early corpus principles
- Corpus linguists are well suited to data curation
 - Continuing need for educating corpus linguists!



The image shows a screenshot of the Merriam-Webster website. At the top, there is a search bar with the text 'virtuous circle' and a magnifying glass icon. Below the search bar, there are navigation tabs: 'Definition', 'Example Sentences', 'Word History', and 'Entr'. The 'Definition' tab is selected and highlighted with a red underline. Below the tabs, the word 'virtuous circle' is displayed in a large, bold, black font, followed by the word 'noun' in a smaller, blue font. Underneath, there is a section for 'variants or virtuous cycle' with a definition: ': a chain of events in which one desirable occurrence leads to another which further promotes the first occurrence and so on resulting in a continuous process of improvement'. The definition text is highlighted in purple.

NOT



Images: www.merriam-webster.com and pixabay.com/

Thank you!

Comments or questions?

Lynne Bowker
(lynne.bowker.1@ulaval.ca)

Canada Research Chair
in Translation, Technologies,
and Society



UNIVERSITÉ
LAVAL