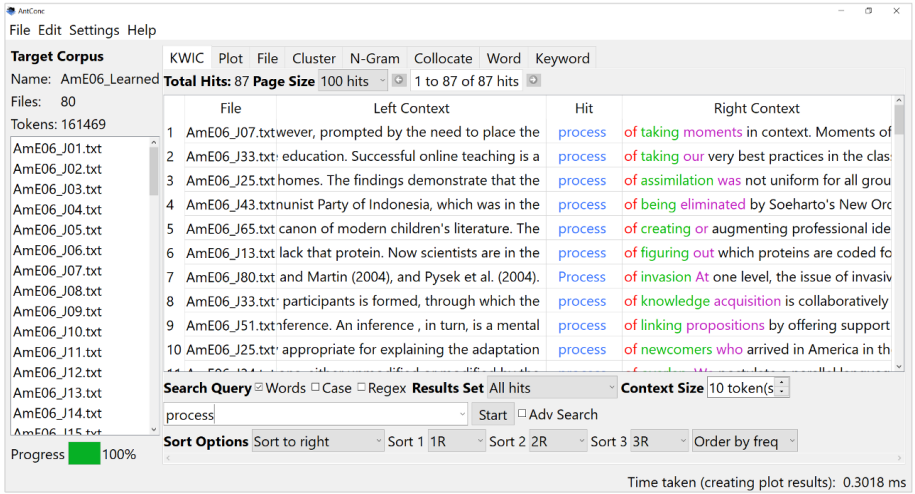
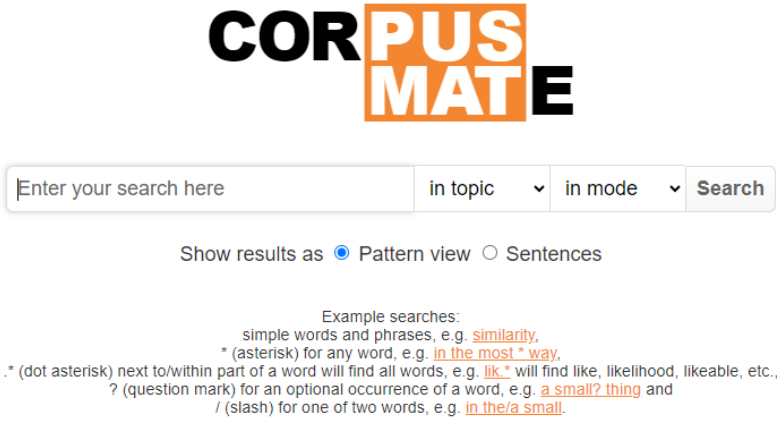
Methodological considerations in the selection and design of corpus tools

Peter Crosthwaite (Univ. of Queensland)

Laurence Anthony (Waseda University)





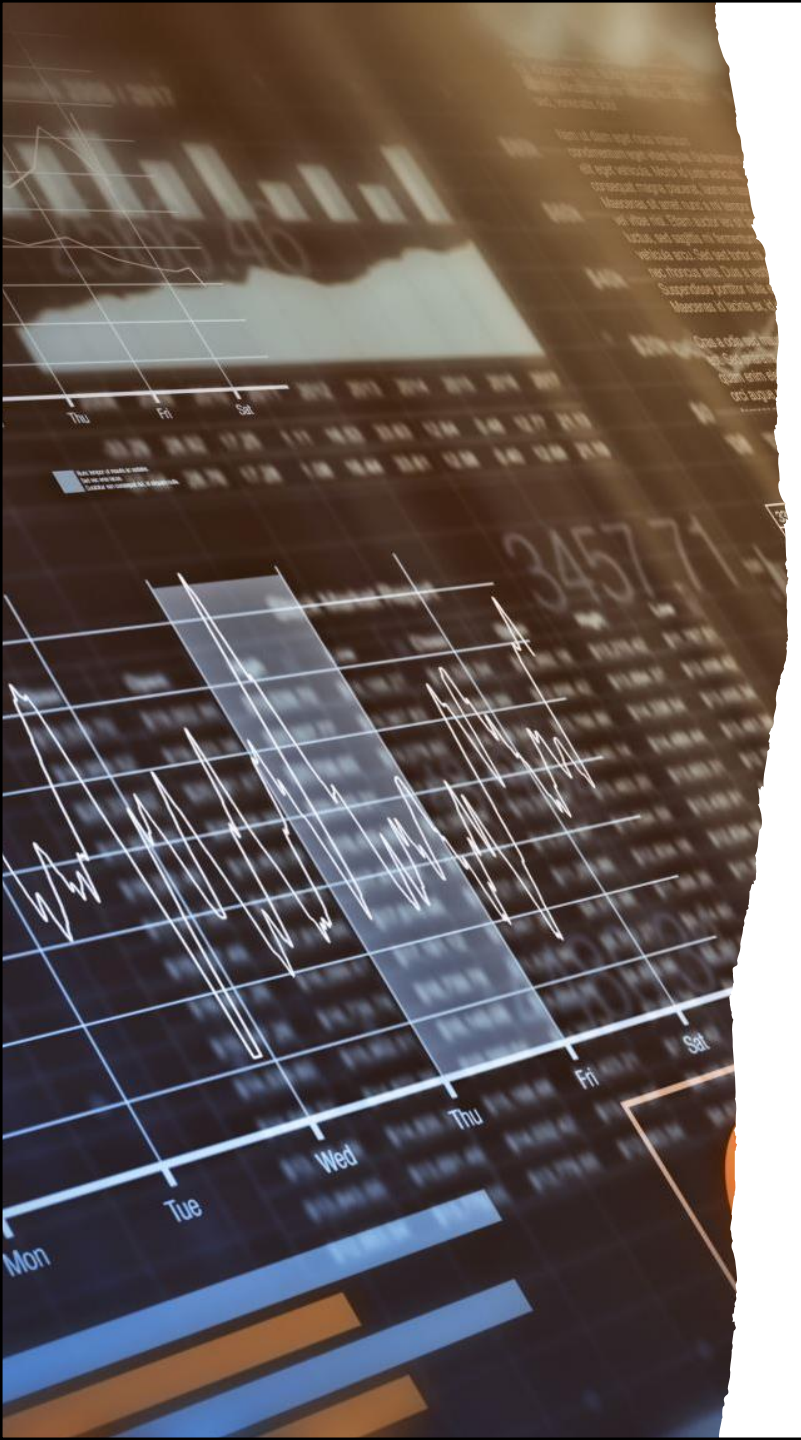
Introduction

Corpus linguists need to make important methodological decisions as they journey through a corpus-based or corpus-driven study.

These decisions include the corpus size, sampling frame, annotation scheme, the choice of tagging, annotation and statistical tools, as well as visualization strategies.

The software tools (web scrapers, taggers, concordancers, etc.) used to assist in building, managing, querying, and analyzing corpora are often presented uncritically.

More transparency in software tool selection and design can support open science, validation, and replication in applied linguistics research.

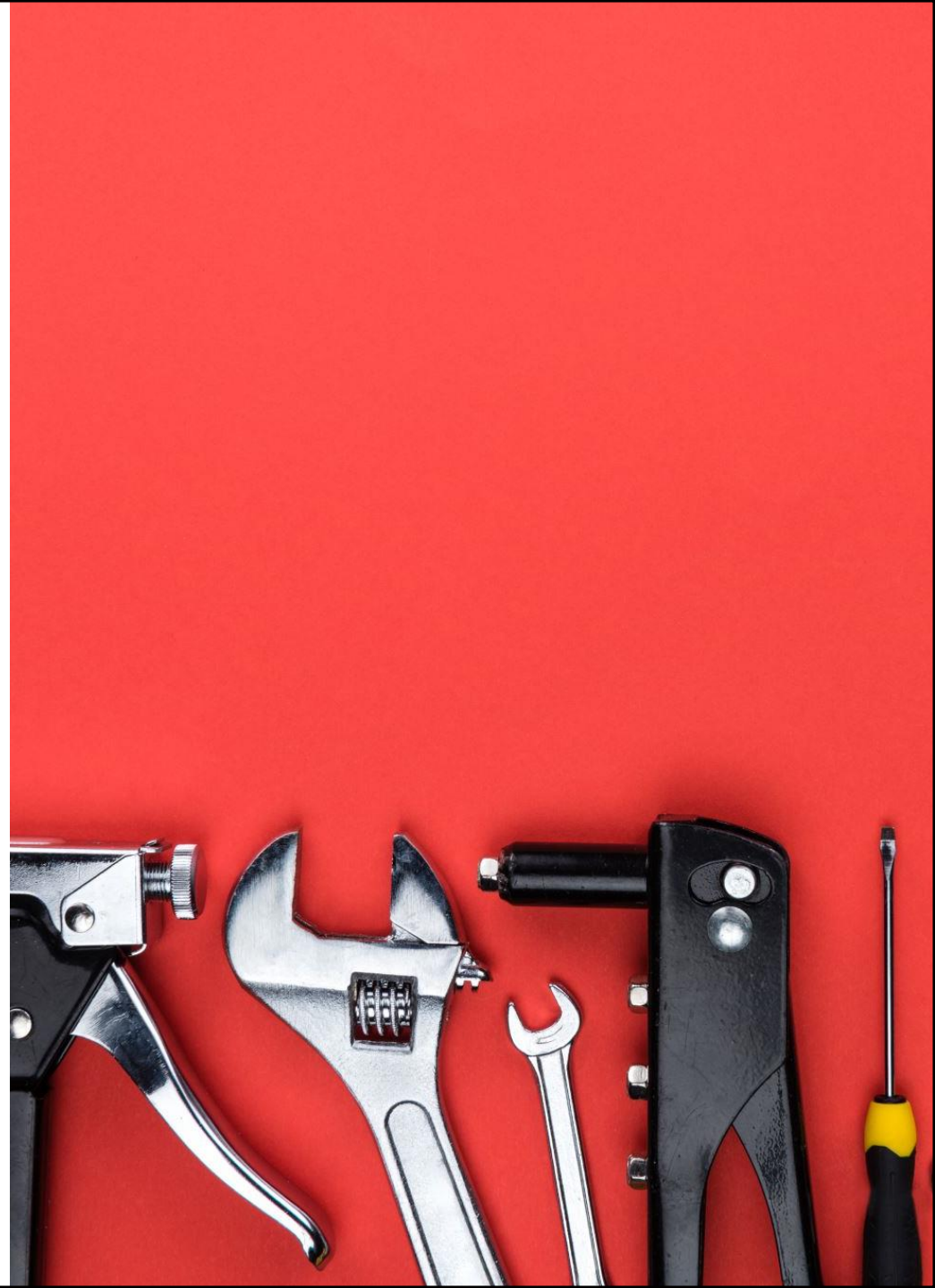


Corpus tools for linguistic research

- Researchers typically use corpus tools to analyse target population language usage using functions like…
  - concordancing
  - cluster/n-gram/lexical bundle analysis
  - collocation analysis
  - word frequency and keyword analysis
- Researchers typically use a single, integrated corpus tool to…
  - simplify data management
  - reduce data interoperability issues
  - enhance analysis by piping results from one function to another
  - facilitate a dynamic QA research process in the discovery process
  - allow for easy validation
- Researchers may adopt DIY corpus tools for…
  - specialized operations and analysis
  - greater speed (?), flexibility (?), transparency, replicability

Corpus tools for linguistic research (general vs DIY)

- General-purpose corpus tools (e.g. AntConc, WordSmith Tools, LancsBox, SketchEngine…
  - are designed for a wide range of corpus sizes and structures
  - include the most common analytical functions
  - include 'best practice' statistical procedures
  - target (more) knowledgeable researchers in the field
  - prioritize data richness over interface simplicity
  - have limited flexibility (stats/visualization options)
- Specialized tools (e.g. R/Python DIY tools)…
  - can meet specific analysis needs
  - can offer better transparency, replicability
  - may offer greater speed and flexibility
  - present challenges in terms of build time, reliability, usage, …

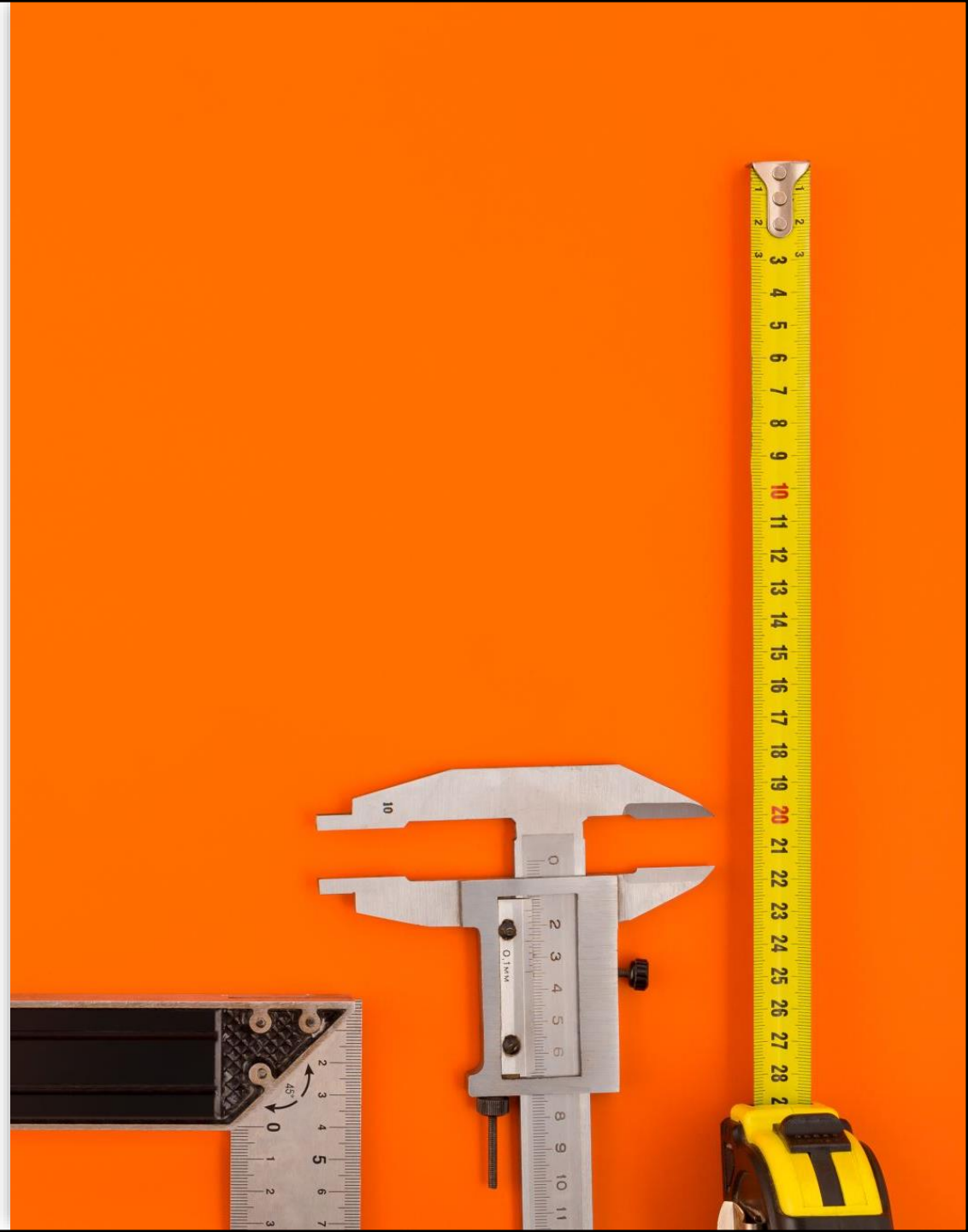


Corpus tools for pedagogic purposes

- Pedagogic (DDL) corpus tools (e.g. SkELL, Linggle, CorpusMate) …
  - target diverse users (teachers and learners) with varying language and computer skills
  - help learners discover language patterns through observation of authentic examples
  - facilitate simple interaction with a corpus, turning learners into investigative linguists
  - prioritize ease of use and 'best practices' for quick learning and analysis.
- Pedagogic (DDL) corpus tools don't usually require…
  - user-upload capabilities
  - complex query options
  - data rich interfaces and statistical options

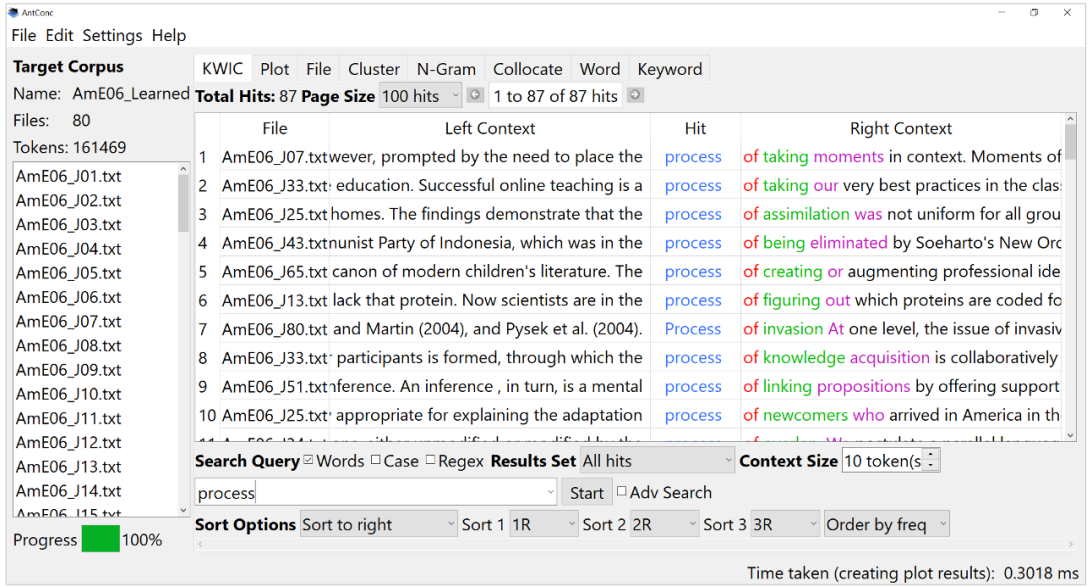
Corpus tools for pedagogic purposes

- Challenges for pedagogic (DDL) corpus tools include…
  - a tendency to use general-purpose tools designed for researchers (with complex interfaces)
  - a lack of knowledge about the needs and preferences of learners in the DDL classroom
- Advancements to pedagogic (DDL) corpus tools include…
  - improvements to general-purpose tools educational use (e.g., the 'KWIC patterns' feature of AntConc).
  - the development of specialized DDL tools (e.g., SkELL and Linggle) that address previous usability challenges.



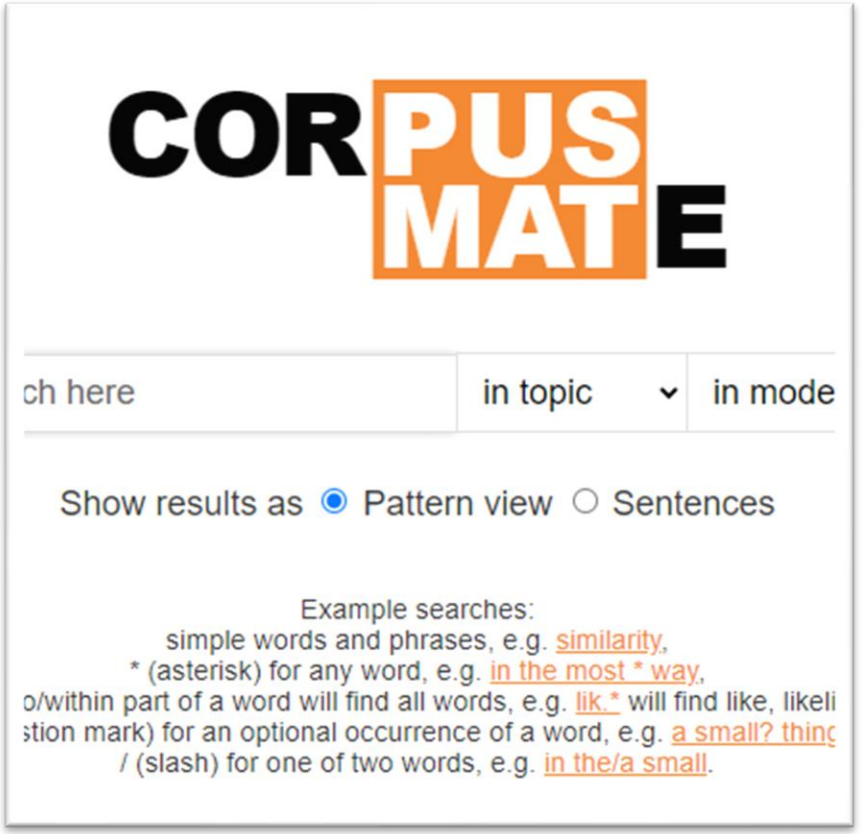
AntConc as an example of a hybrid general purpose/DDL corpus tool

- Design philosophy
  - AntConc provides a scalable learning platform, empowering users to expand their corpus analysis skills as they grow from beginner to advanced levels.
- Main features
  - Started as a simple concordancer function for DDL
  - Evolved into an advanced general-purpose corpus analysis toolkit
  - Offers comprehensive linguistic analysis functions (concordancing, cluster analysis, etc.)
  - Includes features for (advanced?) DDL users (KWIC patterns, word cloud generation, …)



CorpusMate as an example of a dedicated DDL corpus tool

- Design philosophy
  - CorpusMate hopes to overcome barriers to DDL (e.g., system complexity) through the use of a familiar interface like Google's.
- Main features
  - Developed after evaluating existing tools and user feedback
  - Focuses on simplicity and continued use beyond initial interventions.
  - Uses a single search bar, filters for topics, disciplines and registers, plus options for concordance display formats to cater to non-linguists.
  - Includes straightforward guidance on basic query functions and operators to ensure ease of use and to encourage adoption and long-term engagement



Corpus creation and management functions:

Questions to consider – General purpose tools

Does the tool accept the data file type (e.g. Txt, xml, docx, pdf)?

Does the tool handle the data encoding (e.g. UTF-8, UTF-16LE, ...)?

Can the tool work with both left-to-right and right-to-left language data?

What file number and file size limits does the tool have?

What automated processing does the tool provide (e.g. POS tagging)?

Corpus creation and management functions:

Questions to consider – Pedagogical tools

Does the tool have a corpus repository?

Is the repository sufficiently large and diverse to enable the tool's full functionality?

Does the repository have a comprehensive coverage of linguistic variations and contexts?

Are the repository corpora tailored to suit the specific classroom context(s) where the tool is intended for use?

Are the repository corpora tailored to suit the stated pedagogic goals where the tool is intended for use?

Corpus query functions
Questions to consider

Does the tool allow for single and multi-word searches (with/without wildcards)?

Does the tool allow for case sensitive/insensitive searches?

Does the tool allow for POS and lemma-based searches?

Does the tool allow for searches utilizing regular expressions (regex)?

Does the tool allow for searches utilizing the Corpus Query Language (CQL)?

Can the tool be used to search within and/or across sub-corpora, language, dialect, or register?

Corpus query functions:

Link to research/language -related goals

- Is the expression “to be or not to be” found outside of Shakespeare? (word/phrase)
- Is ‘light’ usually used as a noun or adjective? (part of speech)
- How frequently do passive voice constructions appear in scientific research papers? (syntactic construction)
- How frequently do passive voice constructions appear in scientific research papers compared with newspaper articles? (corpus comparison)
- How does the representation of climate change differ in English, French, and Chinese newspapers? (language)
- How does the usage of modal verbs differ across Standard British English, Scottish English and Welsh English dialects? (dialect)
- How frequently do passive voice constructions appear in scientific research papers compared with Standard British English (register)

Concordancing

- KWIC is often the central function in corpus tools designed for general purpose linguistic research.
  - Research question might include:
    - How is the term "artificial intelligence" used in different contexts within academic literature?
    - In what contexts does the word "sustainability" appear in environmental policy documents?
    - In what different ways is the word 'freedom' used in political discourse compared to its use in literature?
- KWIC is often a central component of DDL.
  - DDL creates conditions for learners to notice usage patterns in line with a usage-based account of (second) language acquisition (see Römer, 2023).
    - What tense should I use when writing an abstract?
    - How does Shakespeare use the word 'love'?

	File	Left Context	Hit	Right Context
1	AmE06_J07.txt	wever, prompted by the need to place the	process	of taking moments in context. Moments of
2	AmE06_J33.txt	education. Successful online teaching is a	process	of taking our very best practices in the clas
3	AmE06_J25.txt	homes. The findings demonstrate that the	process	of assimilation was not uniform for all grou
4	AmE06_J43.txt	nunist Party of Indonesia, which was in the	process	of being eliminated by Soeharto's New Orc
5	AmE06_J65.txt	canon of modern children's literature. The	process	of creating or augmenting professional ide
6	AmE06_J13.txt	lack that protein. Now scientists are in the	process	of figuring out which proteins are coded fo
7	AmE06_J80.txt	and Martin (2004), and Pysek et al. (2004).	Process	of invasion At one level, the issue of invasiv
8	AmE06_J33.txt	participants is formed, through which the	process	of knowledge acquisition is collaboratively
9	AmE06_J51.txt	nference. An inference , in turn, is a mental	process	of linking propositions by offering support
10	AmE06_J25.txt	appropriate for explaining the adaptation	process	of newcomers who arrived in America in th

Concordancing:
Questions to consider

- Does the KWIC concordancer offer sorting to the left, center, and right of the search query term?
- Does the KWIC concordancer offer sorting by other measures (e.g., file ordering)?
- Can the results of a KWIC concordancer be ‘thinned’ in any way?
- Does the tool offer a way for the KWIC concordance results to be displayed in ‘KWIC pattern’ frequency order?
- Are the salient patterns in the KWIC concordancer results highlighted in any way (e.g. using colors or labels)?
- Is the surrounding context determined by character count, word count, or sentence boundaries?
- Does the tool offer a way for users to expand the KWIC concordancer to see more surrounding context?

Corpus statistics and data visualisation: Questions to consider

Most tools offer a range of statistical measures for word frequency and dispersion, collocation strength, and keyness, and ways to visualize the results.
(e.g., tables, graphs, word clouds, ...)

Based on these measures, researchers can address questions such as:

- How has the frequency of the expression 'climate emergency' changed in international news media over the past two decades? (frequency and temporal distribution)
- What are the key linguistic differences in the vocabulary used in high school science textbooks compared to university-level science textbooks? (frequency and dispersion)
- What are the most common adjectives used in collocation with the word 'innovation' in corporate annual reports? (collocation strength)
- Which keywords are significantly more prevalent in online health forums discussing mental health compared to general health-related news articles? (keyness).

Corpus statistics and data visualisation
Questions to consider

Does the tool offer the ‘best practice’ statistical measures needed for the investigation?

Does the tool offer older and/or less well-known statistical measures that can be useful for replication and comparison studies?

Can the user modify measures or add additional measures to the tool to keep up with trends in the field?

What visualization methods are available directly in the tool?

Can the results of statistical methods in the tool be easily exported for further processing and visualization using an alternative tool?

Are the statistical methods and visualization methods employed in a tool open to scrutiny?

Corpus statistics and data visualisation:
Pedagogical tools

- Tools need to prioritize simplicity over statistical measures.
- CorpusMate exemplifies this approach (limited statistical functions) promoting engagement.
- CorpusMate presents frequency information and query comparisons in a straightforward, accessible manner.
- CorpusMate visualizes keyness across topics using color-coded line bars to indicate relative significance.
- CorpusMate underscores the importance of selecting appropriate corpus tools and statistical measures.

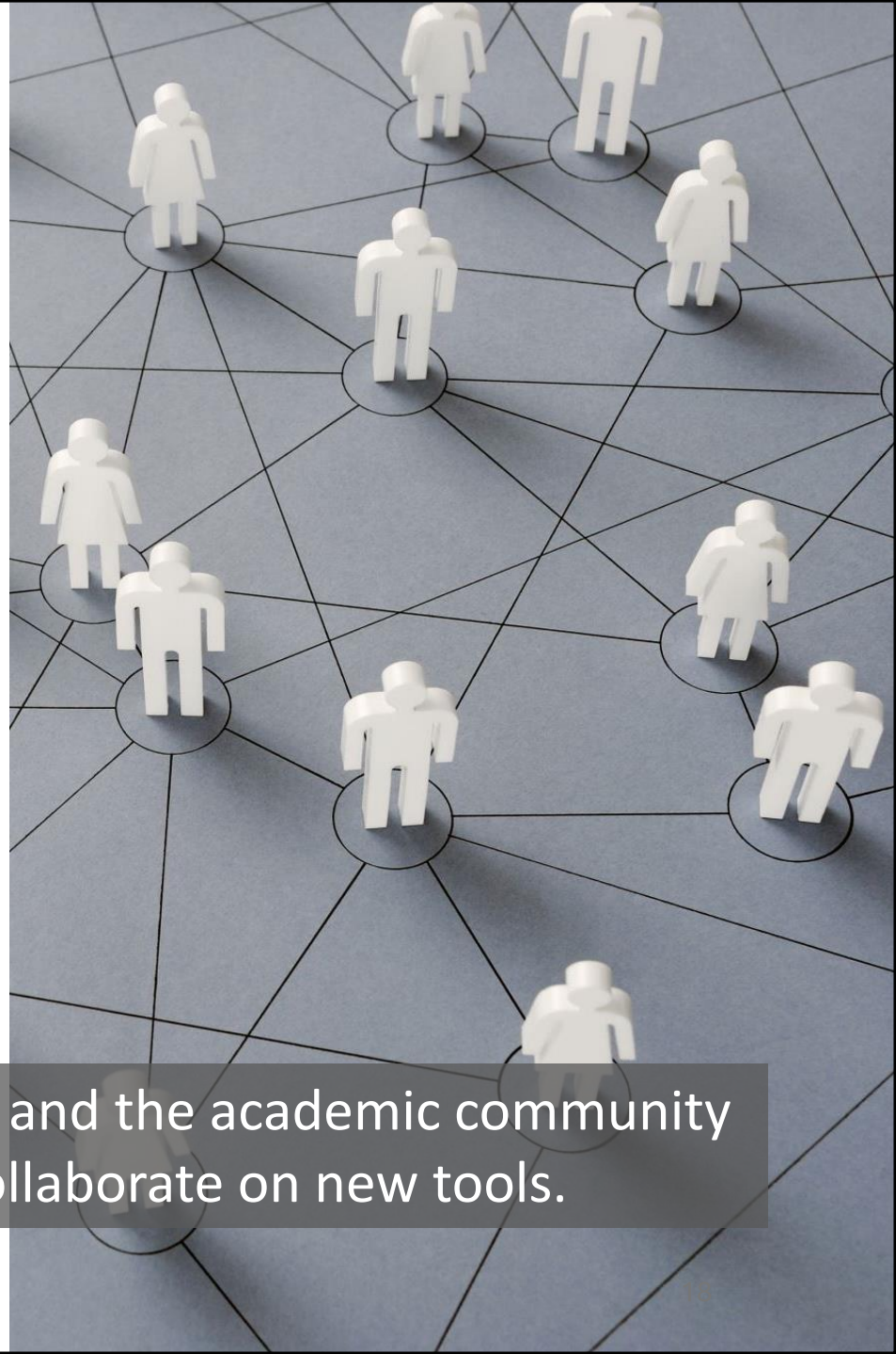
Topic distribution of documents

Topic	Documents	Within topic
Culture, Arts and Music	3,299	31.61%
Society	1,636	27.57%
Business and Economics	1,268	25.3%
Psychology	557	24.21%
Politics	816	21.48%
Technology	1,669	21.32%
Education	665	20.46%
English Language and Literature	755	17.99%
Health and Medicine	1,634	16.56%
Law	398	14.83%
Engineering	347	13.61%
Journalism	83	13.58%
Science	1,904	13.14%
Geography, Agriculture and Environment	779	13.05%
Physics	765	12.12%
History	1,279	11.53%
Biology	1,645	11.35%
Architecture, Planning and Design	227	8.38%
Chemistry	263	6.58%
Mathematics	379	4.11%

Click a topic to see pattern/sentence results for that topic.
Blue horizontal bar: the search result is more likely to be used in that topic compared with the average across all topics. Red bar: the search result is less likely to be used in that topic compared with the average across all topics. The longer the bar the more/less likely your search result is likely to be used in that topic compared with the average across all topics.

Future directions
Collaboration and community engagement

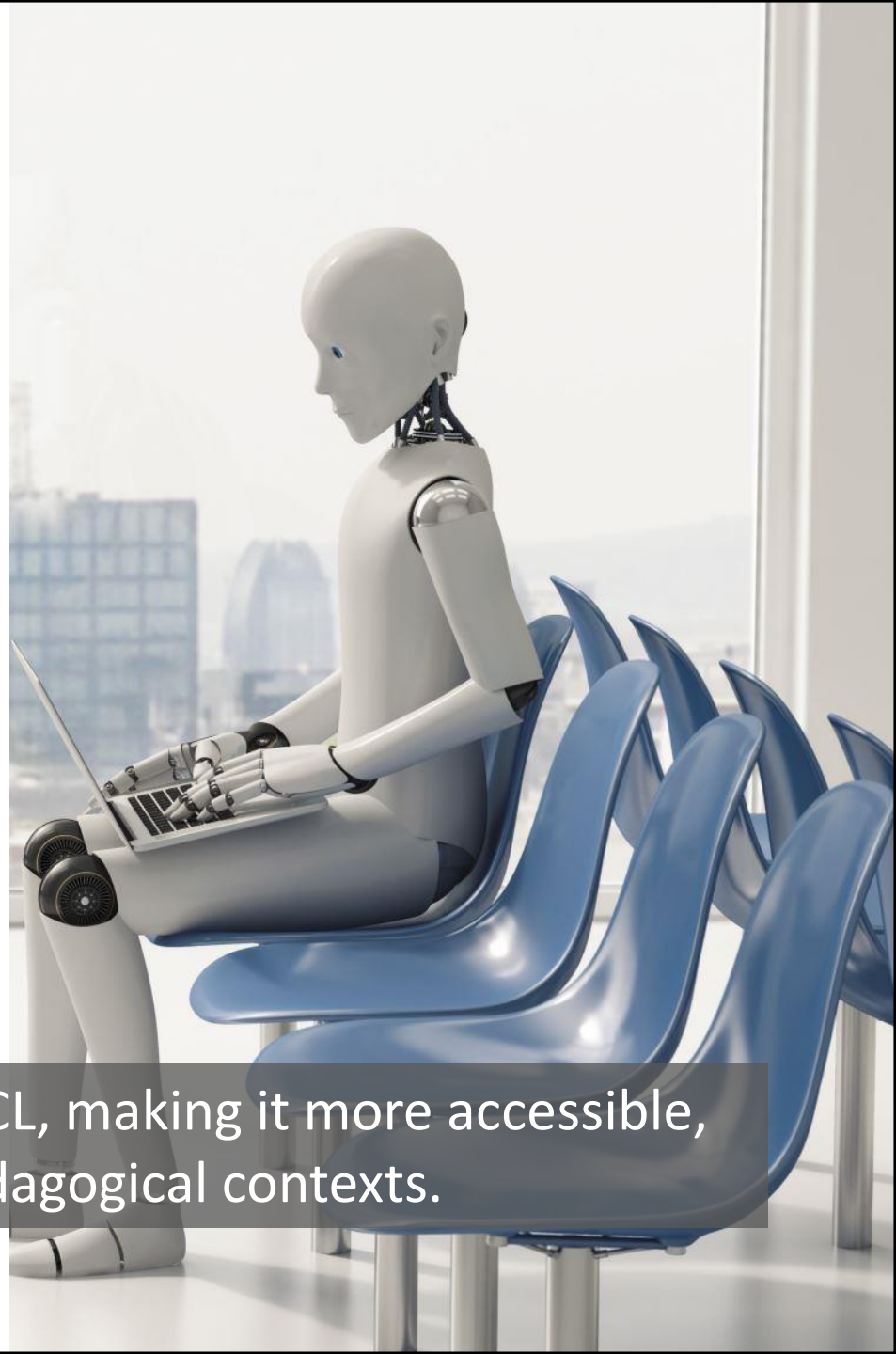
- The post-COVID era has seen an increase in online interconnectivity among corpus and applied linguistics researchers (e.g., social media groups for general and specific interests like DDL.)
- Disciplinary associations and groups are now widespread, featuring thousands of members globally.
- Discussions on corpus tool usage and selection justification remain less common (in general forums).



There is a need for a unified forum for tool developers and the academic community to compare tools, discuss developments, and collaborate on new tools.

Future directions
Integration of CL with AI

- LLMs are significantly impacting language research, teaching, and learning (e.g., tagging, annotation, analysis)
- Technical challenges exist for integrating CL with AI (e.g., hallucinations, cost, replicability)
    - AntConc's integration of a "ChatAI" feature will enable users to combine traditional corpus analyses with LLM insights through natural language queries.



Future AI-assisted CL tools are likely to revolutionize CL, making it more accessible, more nuanced, and more applicable in pedagogical contexts.

Q&A