How humans vs. machines identify discourse topics: an exploratory triangulation

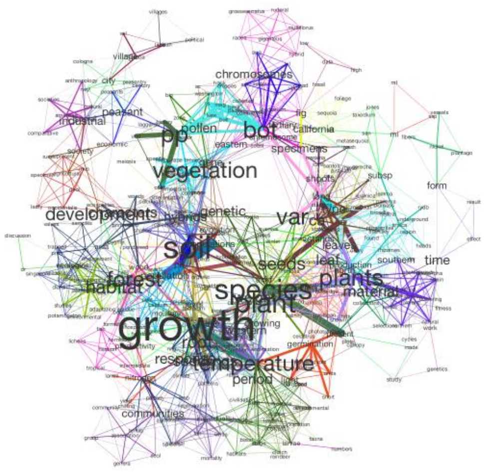
Sylvia Jaworska & Mathew Gillings

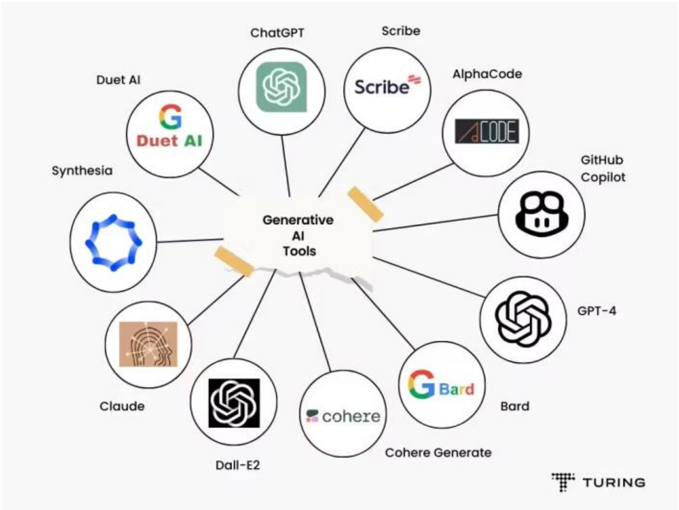
University of Reading

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

Purpose

- Exploratory triangulation using a quasi-experimental approach
- Understand what linguistic and non-linguistic methods can and can't do for researchers interested in discourse and discourse topics
- Contribute to the growing line of inquiry in corpus linguistics (Marchi & Taylor, 2009; Baker & Egbert, 2016; Egbert & Baker, 2020; Gillings and Hardie, 2022; Curry et al.; 2023)
- Interdisciplinary research and exposure to different methods and tools, e.g. NLP, sentiment analysis, topic modelling, vector analysis, word embeddings etc.
- New tools: generative AI based on LLMs but the problem of explainability





Overview

- Methods in Textual Analysis
- Methods and Conditions under investigation
    A. Topic modelling with topics labelled by ChatGPT (machine + no context + machine)
    B. Topic modelling with topics labelled by eyeballing (machine + no context + human)
    C. Concordance analysis (machine + more context + human)
    D. Close reading (context + human)
- Our Corpus & Quasi-Experimental Approach
- Results
- Preliminary Conclusions

Discourse topics

- Different names and conceptualisations: sentence topics, discourse topics, topics, global proposition, subject, aboutness (Watson Todd, 2016)
- “whatever it is that is being talked about” (Brown and Yule 1983b, 62)
- Aboutness (Scott, 2006) fits well our purpose; cline of aboutness from no aboutness to minor aboutness to great aboutness (text summary).
- Important for discourse analysts and social scientists working with/on discourse.
- Significance, cohesion, connectedness, focus, foregrounding backgrounding etc.

Approaches to topic identification

Qualitative	Mixed	Quantitative
• slow • intuitive, interpretative, qualitative, hermeneutic, qualitative with some quantifications (e.g., content analysis); • based on a small data sets; • issues around cherry-picking, representativeness, time, etc but control over the data set and procedures.	Bridge quantitative and qualitative divide: CADS	• fast • quantitative, computational, text mining, machine/deep learning, gen AI (LLM-based); • huge data sets • powerful algorithms but difficult to know how they actually produce the results they do; users have less or no control over the data and procedures; • need for empirical validations to understand what kind of job they can do for us; testing how these tools and methods work in practice.
How do they compare in terms of the outputs that they produce?		

Triangulation

- Marchi and Taylor (2009) draw upon Denzin (1970) to identify 4 types of triangulation:
  - **Investigator triangulation**: using more than one researcher to explore the data.
  - **Data triangulation**: collecting data through several sampling strategies.
  - **Theoretical triangulation**: exploring data through more than one theoretical lens.
  - **Methodological triangulation**: using more than one method to collect and analyse data; also referred to as between-method triangulation.
  - CADS, then, is already 'triangulated'.

Triangulation

Baker and Egbert (2016)

- Same dataset given to different researchers
- Researchers asked to perform different CL techniques on the data: keyword analysis, collocation analysis, multi-dimensional analysis, etc.
- Comparison of CL approaches in the final chapter; only a few cases of shared findings; "most of them [corpus methods] will lead to different parts of Rome" (p. 207)

Egbert and Baker (2020)

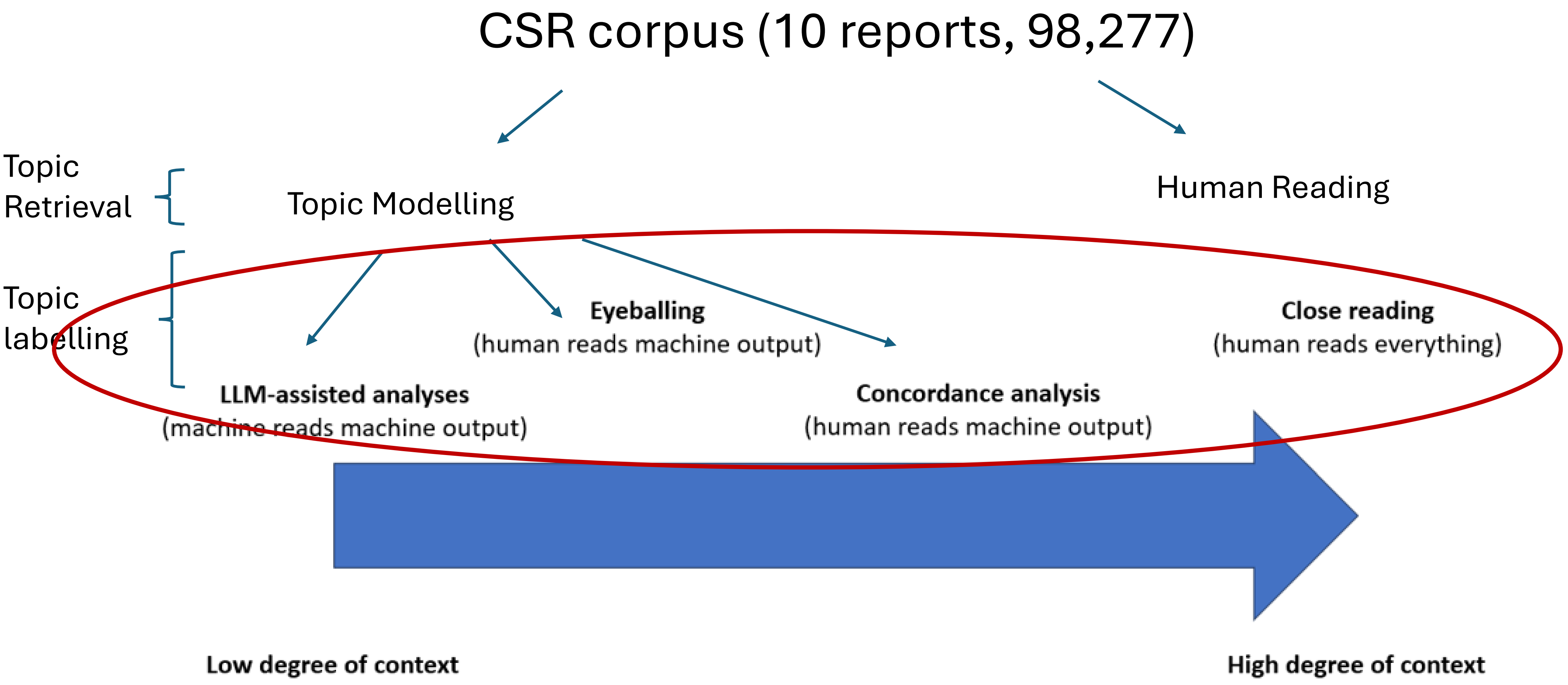
- Different chapters with scholars combining CL with another method from within linguistics
- Other methods included psycholinguistic analysis, pragmatic analysis, experimental etc.
- Showcases different methodological choices available to the researcher
- Increased ecological validity, research collaborations and potential for synergies
- Ways of triangulating CL methods with other methods

    (1) Convergent vs Correlational

    (2) Independent vs Sequential vs Cyclical

Our approach: combining CL with a method from *outside* linguistics; convergent and independent

Our study


CSR corpus (10 reports, 98,277)
Topic Retrieval
Topic Modelling
Human Reading
Topic labelling
Eyeballing
(human reads machine output)
Close reading
(human reads everything)
LLM-assisted analyses
(machine reads machine output)
Concordance analysis
(human reads machine output)
Low degree of context
High degree of context

Corpus under analysis

- Sustainability reports from 2021; removed letters from CEO, appendices, graphs, etc.
- Sustainability an important matter: 64% of adults in the UK worried or very worried about climate change (UK's Office for National Statistics, 2023)
- Written for wider audiences and stakeholder groups, of which members of the public are (supposed) to be the most important. The language of such reports is therefore less technical, and they are written in a way that should be accessible to average adult readers.

Industry	Company	Number of words
Pharmaceutical	AstraZeneca	11,624
	GlaxoSmithKline	8,611
Food	Arla	13,292
	Nestle	17,563
Oil	Exxon	10,070
	CNNOC	2,932
Banking	Lloyds	10,430
	Santander	4,193
Manufacturing	Apple	12,877
	Ikea	6,685
Total number of words:		**98,277**

Topic modelling

- Machine learning algorithm with roots in computer science, but applied to and used within the digital humanities
- The user inputs a large corpus of texts, and the LDA algorithm detects patterns of co-occurring words across texts; these sets of co-occurring words are ‘topics’
- Researcher can control parameters: number of topics to be found, include/exclude stopwords, number of sampling iterations, etc.
- Two outputs presented to the researcher:
    - List of topics (represented by 10 lists of 10 co-occurring words)
    - Composition document showing the distribution of those 10 topics across the texts

- It’s the researcher’s task to determine a topic label for each list of words, which is typically achieved via eyeballing (Gillings and Hardie, 2022).
- We used the Machine Learning for Language Toolkit (MALLET; McCallum, 2002), the most widely-used LDA program for digital humanities. We opted to exclude stopwords and asked the tool to identify 10 topics consisting of 10 words each.

Topic modelling output

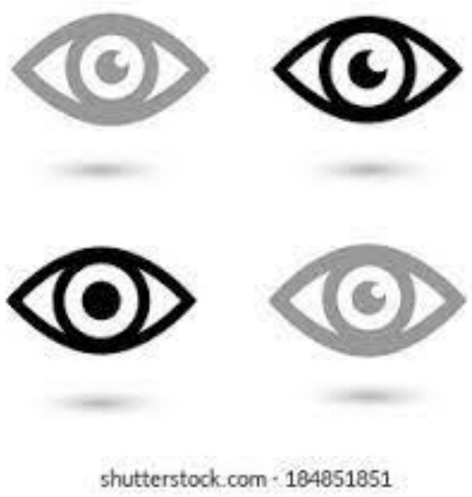
Topic number	Words making up that topic
1	*food nestlé water business systems approach forest supply regenerative agriculture*
2	*ikea energy products materials renewable product recycled emissions climate chain*
3	*climate business including working training development improve sustainability focus supporting*
4	*health healthcare water patients programme medicines clinical patient data systems*
5	*work impact local communities access reduce solutions make part environmental*
6	*support key provide million risks management future year natural employee*
7	*global people rights human health sustainable products emissions operations employees*
8	*board exxonmobil company gas energy management waste plastic employees development*
9	*arla dairy farming food milk consumers waste carbon owners sweden*
10	*colleagues customers group support financial santander health digital programme skills*

Method A: LLM-assisted topic labelling

- Since late 2022, large language models (LLM) have been cast into the public consciousness, increasingly being used in workplaces to aid in the completion of everyday tasks. On its most basic level, an LLM is a text prediction system, which uses deep learning algorithms and the vast Internet data (fine tuning based on human annotators).
- One such LLM, ChatGPT, is “an artificial intelligence (AI) chatbot that processes and generates natural language text, offering human-like responses to a wide range of questions and prompts” (Doshi et al., 2023: 6). ChatGPT generates predicted text, based on user-provided prompts (G - generative, predicts the next word; P - pre-trained; T – transformer).
- Curry et al. (2023) used ChatGPT to perform (part of) the analysis, in place of traditional corpus methods. They found that ChatGPT was reasonably effective at semantically categorising keywords and assigning a category label; yet poor at performing concordance analysis, and poor at form-to-function analysis.
- We take the topic model output and ask ChatGPT to assign a topic label to them. **2 analysts worked independently but** each used the exact same prompts to ask ChatGPT for a topic label: “Look at the words and tell me what the overarching topic is. [list of 10 words].” This was repeated 10 times, once per topic.

Method B: Topic modelling + eyeballing

- It's the researcher's task to determine a topic label for each list of words, which is typically achieved via eyeballing (Gillings and Hardie, 2022).
- **2 different analysts** were given the 10 topics (of 10 words each) and based purely on their knowledge of the discourse and wider world knowledge, asked to assign topic labels. They worked independently.



Method C: Concordance Analysis

- Method C takes the topic modelling output (the 10 lists of 10 co-occurring words) and then asks analysts to assign labels based not via eyeballing, but via concordance analysis (Gillings and Mautner, 2023).
- We uploaded the corpus to Sketch Engine and **2 different analysts** were asked to run a concordance analysis for each of the words that appeared in the topic model (100 words in total). They used a random sample of 100 concordance lines and, based on their reading through those lines, assigned topic labels.

Left context	KWIC	Right context
olleagues, companies, and advocates to further our efforts to make our environmental work a force for good in	people's	lives–and give the communities most impacted by climate change a seat at the table. As a result, this has beei
noved ahead with greater urgency than ever before to create a stronger, healthier future for our planet and her	people	. In 2020, that meant real progress in our fight against climate change. Apple became carbon neutral for our w
proof that the fight against climate change is also a fight for local economies, for the rights of indigenous	peoples	, and for the communities whose lives and livelihoods are most threatened by climate change. These are syste
we do have are goals to strive for, and a global community of businesses committed to doing the right thing by	people	and the planet. Thank you for your role in our progress to push this urgent work forward. Lisa Jackson Vice Pr
inclusive opportunities. Our approach Climate change is one of the greatest threats of our time, putting at risk	people's	access to clean air, adequate food, safe drinking water, and sanitation. This means the impact of the changes
challenge is significant. But so is our potential to have an impact. The changes that we push forward affect the	people	who interact with our products, influence the markets in which we operate, and create change for broader glob
reasingly scarce and vulnerable to the effects of climate change. As a community resource, water is shared by	people	and ecosystems across very different environments. Our efforts to reduce our freshwater withdrawals and retu
nd HVAC systems across our facilities on an ongoing basis, to adapt to reduced capacity and use patterns. As	people	return to our facilities, we're constantly monitoring occupancy levels to determine what additional ventilation
gins with the substances in our products. By focusing on safety in our product designs, we strive to protect the	people	who design, make, use, and recycle our devices. Through close engagement with leading members of the scie
of our owners, the continuity of services and operations for our customers and the health and safety of our	people	. Adapting to recurring lockdowns, re-openings, labour and supply challenges and inflation has become an inte
to grow, not least because of milk's rich content of protein and calcium, versatility and affordability. To ensure	people	continue to trust and enjoy dairy benefits, we and our farmer owners are committed to leading the way on sust
Arlagården® includes four focus areas; milk quality and food safety, animal welfare, climate and nature, and	people	. To ensure compliance, farmer owners are audited by an external certification body. Each year, 30 per cent of
ESS TO HEALTHy NUTriTiON – fOOD SAfETy We provide safe, healthy and affordable dairy products to help	people	eat healthily and sustainably around the world. By balancing the environmental impact of production with the
l, we strive to contribute to the realisation of the UN Sustainable Development Goals. Food safety first Helping	people	to eat a healthy diet is important, but first and foremost we must make sure that our products are safe to eat ai

Method D: Close reading of texts

- No machine assistance
- **2 different analysts** read independently through all 10 texts and were asked to firstly, identify 5 key topics for each individual text, and then secondly, based on that, they identified 10 key topics for the entire corpus. Thirdly, they were asked to list the 10 most salient words for each topic.
- Our interest was in the 10 key topics that analysts decided on for the entire corpus.

Two sets of results

Comparing analysts' responses *within the same* condition:

allows us to assess inter-rater reliability for a method

Comparing analysts' responses *across* conditions:

allows us to assess how different methods pick up on different topics

Given the same data, method, and instructions, do different analysts arrive at the same outcomes? Probably not, but where precisely lie the differences and similarities, to what extent and possibly why?

Method A: ChatGPT

Topic number	Analyst 1	Analyst 2	Similarity score
1	Sustainable Agriculture and Environmental Stewardship in the Food Industry	Sustainable Practices in the Food and Beverage Industry	2
2	Sustainable Practices in IKEA's Supply Chain and Product Development	Sustainability in the Retail Industry	2
3	Business Sustainability and Climate Focus through Training and Development	Business Sustainability and Climate Focus	1
4	Healthcare Systems and Patient Data Management	Healthcare Systems and Patient Data Management	1
5	Environmental Solutions for Local Community Impact	Environmental and Social Impact Solutions	1
6	Employee Support in Natural Risk Management for Future Sustainability	Risk Management and Employee Support for Future Sustainability	1
7	Global Sustainability and Human Rights in Corporate Operations	Global Human Rights and Sustainable Operations	1
8	ExxonMobil: Energy Management and Sustainable Development	Energy and Environmental Management in ExxonMobil Company	1
9	Sustainable Dairy Farming and Food Practices: Arla in Sweden	Sustainability in Arla Dairy Farming and Food Production	1
10	Santander: Financial Support and Digital Health Skills Program for Colleagues and Customers	Santander's Financial and Health Support Program for Colleagues and Customers	1
Mean similarity score:			1.2

(1 = The labels are exactly or almost the same; 2 = The labels have some degree of similarity; 3 = The labels are quite or completely different).

Method B: Eyeballing

Topic number	Analyst 3	Analyst 4	Similarity score
1	Sustainability pipelines and processes	Regenerative agriculture production	2
2	Renewable materials	Sustainable products	2
3	Educating employees about sustainability	Training for sustainability	1
4	Clinical healthcare systems	Healthcare	1
5	*Environmental* impact on local communities	Impact on local communities	1
6	Risk mitigation	Risk management	1
7	Sustainable workplaces	Human rights	3
8	Sustainability management	Corporate management around sustainability	1
9	Carbon reduction	Impact of Swedish dairy farming	3
10	*Education programme*	*Corporate training*	2
Mean similarity score:			1.7

Method C: Concordance analysis

Topic number	Analysis 5	Analysis 6	Similarity score
1	Corporate ethics	Corporate practices	3
2	Responsible sourcing of materials	Sustainable offerings	3
3	Employee career development	Employee career and product development	1
4	Advancements in healthcare	Prioritising health and medicine	1
5	Business/community integration	Sustaining communities and the environment	2
6	Business prosperity	Securing our future: supporting people and planet	3
7	Global impact	Collaborative protection for every individual	3
8	Environment and climate change	Corporate ethics	3
9	Advancements in food and farming	Cultivating sustainability: regenerative farming	2
10	Financial support and wellbeing	Employee and customer support	1
Mean similarity score:			2.2

Method D: Close reading

Analyst 7		Analyst 8		
Topic label	**10 words**	**Topic label**	**10 words**	**Similarity score**
Environment	Forest, agriculture, water, biodiversity, animal welfare, waste, organic, reforestation, natural resources, stewardship	Environment and nature	sustainability, climate, footprint, welfare, biodiversity, ecosystem(s), protecting, resources, health, water	1
Carbon	Emission, footprint, fossil fuels, transition, carbon neutral, netzero, scope, offsetting, greenhouse gas, reduction	Inclusive, carbon-neutral economy	neutral(ity), net-zero, footprint, reduction, renewable, alternative, offsetting, health(y), emissions, forests(s)	2
Climate change	Climate, challenge, action, protection, renewables, climate risk, mitigation, adaptation, resilience, policy	Tackling climate change	Address, complex, problem/challenge/impacts, fight(ing)/combat/tackle, target, value chain, risk(s), health, biodiversity loss, forests/water	2
Products/services	Innovation, R&D, circular, design, smart, packaging, sourcing, life cycle, longevity	Product sustainability, affordability and availability	Sustainable, affordable, available, inclusive, circular, safe, developing, healthy/nutritional, long-lasting/durable, product life cycle	2
Governance	Ethics, conduct, values, audit, transparency, reporting, compliance, regulations, fairness, accountability	Corporate governance	Board, strategic, growth, strategic, (company) value, transparency, culture, sustainability, risk, climate	1
People	Employees, customers, colleagues, suppliers, training, human, rights, recruitment, retention, personnel development, patients	Employee empowerment, development, and engagement	Empowering, development, engagement, retain, training, opportunities, health, safety, care, promoting	2
Health	Nutrition, safety, healthcare, wellbeing, medical insurance, mental health, covid-19, emergency, illness, pandemic	Health	Employee(s)/workers/workforce, system, strategy, mental, physical, public, global/human, local, soil, diet	1
Diversity	Inclusion, ethnicity, gender, sexual orientation, disability, women, LGBTQ+, race, equity, equal opportunities	Diversity and inclusion	Staff, clinical trials, communities, policy, framework, recruitment, gender, ethnic, promote/foster, belonging	1
Company effort	Commitment, support, contribution, goal, target, approach, strategy, partnership, ambition, help	Business ethics	Fair, transparent/transparency, responsible, complying/compliance, values, trust, integrity, conduct, regulate/regulations, legal/law	3
Money	Assets, investment, affordability, pricing, market performance, shareholder value, growth, capital, costs, pay	Safety	Health, well-being, culture, risks, personnel/workforce/employees/patient, manage(ment), environment(al), product, quality	3
Mean similarity score:				1.8

Preliminary observations

- Hardly any textual context available to the researcher in Method B, whereas there is plenty of textual context available to them in Method D. Yet, the similarity metric is uncannily similar: 1.7 in Method B, and 1.8 in Method D.
- Method A’s similarity score of 1.2 means that ChatgPT performed ‘best’ out of all four methods.
- Regardless of whether automated LLM-assistance, eyeballing or close reading is employed, the similarity score between analysts is likely to be similar. Whilst this has no bearing on the quality of the analysis (and thus the ‘true’ discourses being represented), it is an important implication for inter-researcher reliability in that when analysts receive the same instructions to identify topics in the same data set, they are likely to arrive at similar conclusions at the two ends of the spectrum.
- Method C should, theoretically, be the middle-ground in terms of the amount of context available to the analyst, and thus theoretically we might expect a similarity score in the middle too. Yet that is not the case, and the similarity score is 2.1: a slightly higher degree of divergence, in comparison. This appears to suggest that when the analyst is given additional tools and a little more context to help their interpretation, the door is open for increased divergence of opinion.

Comparing responses across conditions

- Responses gathered via Method A differed from all other responses/analyses; they differed in length, focus/vagueness and mixing up domains, e.g.

  Method A: *Risk Management and Employee Support for Future Sustainability*

  Method B: *Risk Management/ Risk Mitigation*

- **Only two topics** were shared across all four methods and all eight analyses: *Healthcare* and *People*; the most salient topics in the corpus.
- There is one further topic that was identified via all four methods, but not necessarily by all eight analyses: *Ethics*. Both ChatGPT outputs had reference to *Human Rights*; one of the Method B analysts identified *Human rights* as a topic; a Method C analyst identified *Corporate ethics*; and a Method D analyst identified *Business ethics*.
- Four topics that were shared across Methods B and C: *Renewable and sustainable materials*, *Employee training*, *Business/community integration*, and *Sustainable farming*.
- Method D analysts were able to identify topics that were nowhere to be found via the other methods. These include *Diversity* and *Diversity and Inclusion*.
- Difficult to identify more as the topics and phrasing was so varied.

I 've studied now Philosophy
And Jurisprudence, Medicine, -
And even, alas! Theology,
From end to end, with labor keen;
And here, poor fool! with all my lore
I stand, no wiser than before.

Faust, J. W. Goethe

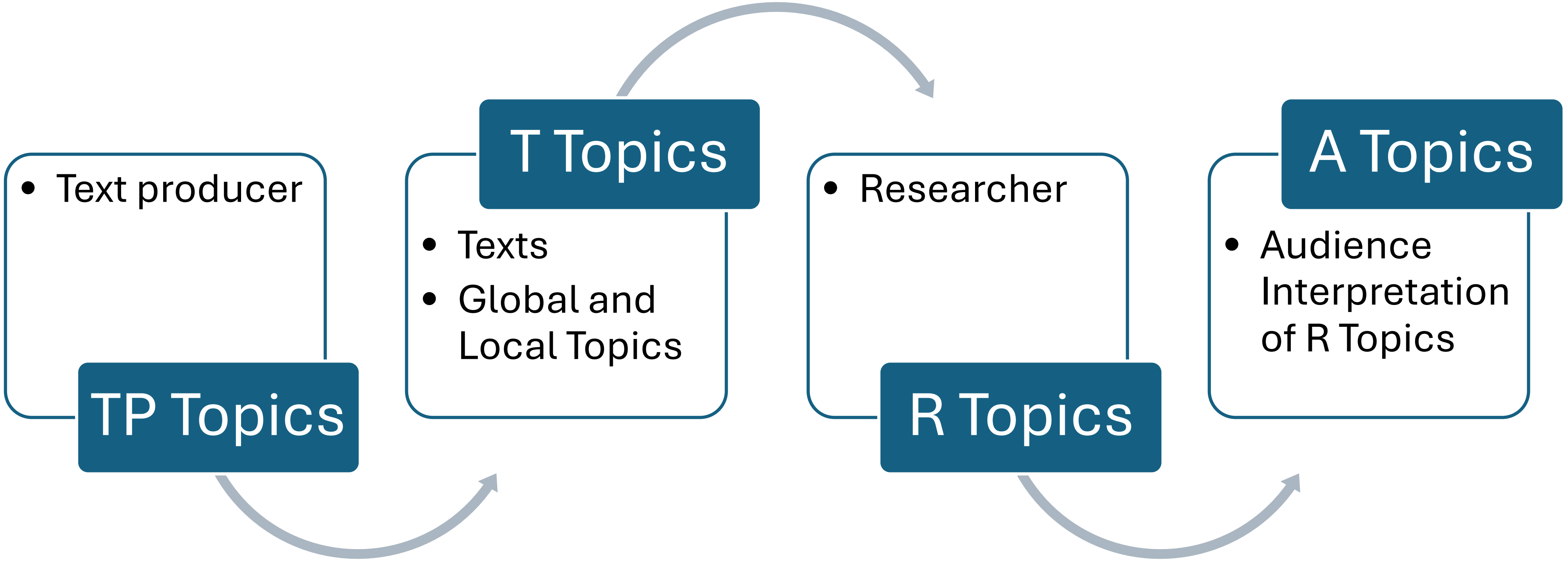
Observations

- What is the gold standard?
- Is Method D the ‘best’ approach because our analysts read the whole corpus and identified topics not accounted for by the other methods?
- Shall we now skip CL tools and use Gen AI to summarise texts because it is pretty consistent?

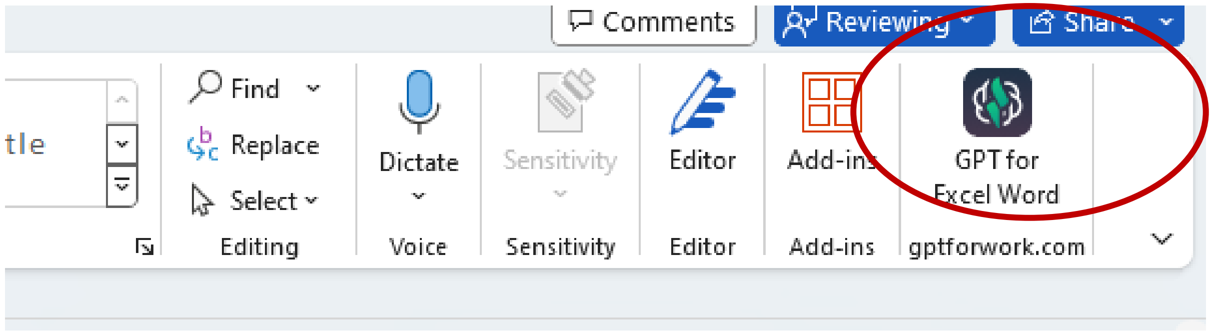
- When humans assign topic labels (Methods B, C and D), they attempt to make each topic as distinct as possible. They seemingly try to label topics in such a way that the reader can imagine different discourses being built up within them. They also tend to be concise and use one- or two-noun labels and thus creating **concepts**.
- Frequency is the basis of machine-assisted approaches, but what is frequent cannot always be assigned significance. It is not always the frequency of the words but the diversity of meanings that different frequent and infrequent words produce in combination. CL has more checks and balances built into the system.
- For the human analysts, frequency was not equal to importance; even if something important is mentioned relatively infrequently, human readers will pick up on it whereas the computer/machine-assisted approaches will not. But is it relevant to the text producer and how they intended their texts to be interpreted?

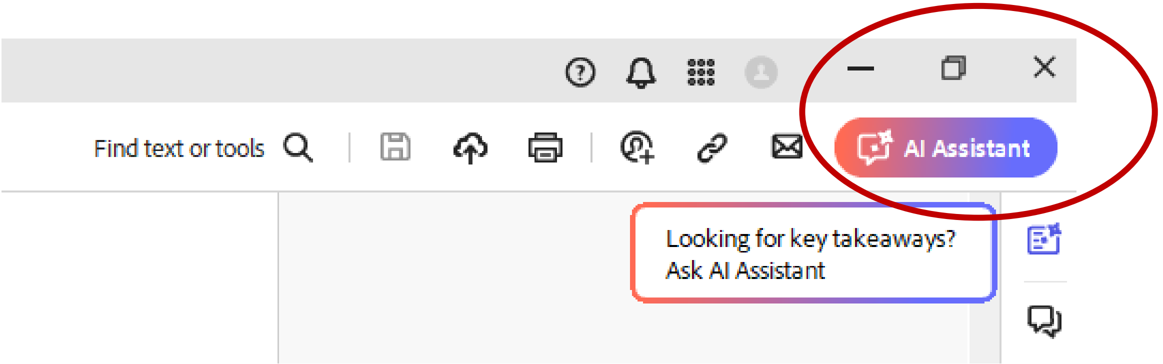
Observations

- Highly decontextualised approaches (i.e., LLM-assisted analyses and topic modelling) and highly contextualised approaches (i.e., close reading) all produce high levels of inter-analyst agreement.
- Interestingly, it was Method C (i.e., concordance analysis) where uncertainty and differences in labelling were most widespread. This is at odds with work in corpus linguistics which suggests that having several analysts working on the same dataset helps to ensure as many discourses are identified as possible (both salient and hidden alike).
- Key point: Concordance analysis points to **local manifestations of topics** that then potentially build onto more global topics. There is therefore the need to validate this and to read parts of the corpus to understand how the local topics construct more global messages.
- Transparency and reflexivity; these principles seem more important than ever especially given the lack of transparency regarding LLMs and the tools that use the models.
- Explainable AI: There is a need to empirically test and validate how the various tools and methods work in practice, and what they can do for us, also in comparison with more established linguistic approaches.
- Methodological protocols with details regarding the outputs and how they were obtained.
- More nuanced understanding and definitions of discourse topics including the different stages of text production, interpretation and reception

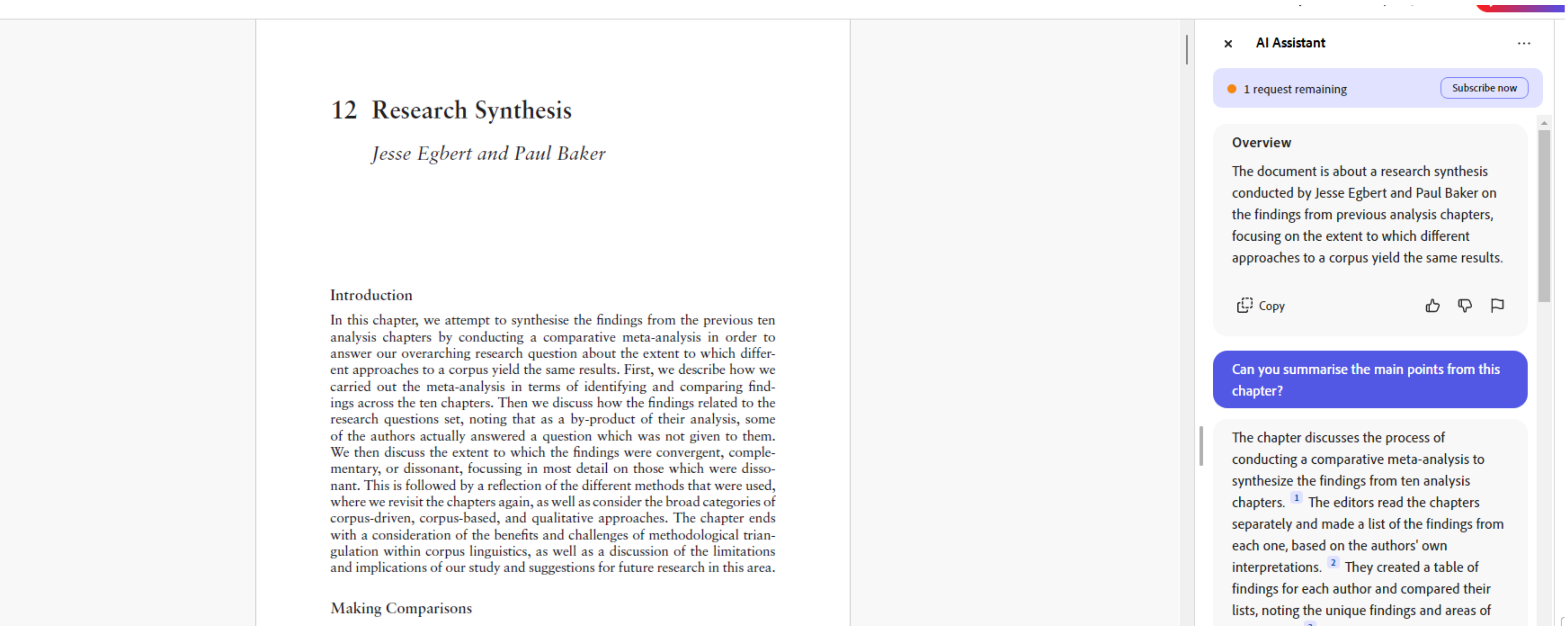

Text producer
TP Topics
T Topics
Texts
Global and Local Topics
Researcher
R Topics
A Topics
Audience Interpretation of R Topics

PS: to add to the mix





PS: to add to the mix



PS: to add to the mix

- The chapter discusses the process of conducting a comparative meta-analysis to synthesize the findings from ten analysis chapters. The editors read the chapters separately and made a list of the findings from each one, based on the authors' own interpretations.
- They created a table of findings for each author and compared their lists, noting the unique findings and areas of agreement. The findings were organized into three sections: those relating to the corpus as a whole, those comparing the four varieties of English, and those comparing the three topic areas.
- The results showed that most of the authors made unique discoveries, with some areas of shared focus and slightly more agreement than disagreement. The findings related to the corpus as a whole included characteristics of spoken discourse, subjectifing strategies in question words, and the use of stance bundles.
- The chapter also discusses the different methods used in the study, including keyword analysis, semantic annotation, lexical bundles, multi-dimensional analysis, collocation networks, variationist analysis, pragmatic features, gendered discourses, qualitative analysis of stance, and stylistic perception analysis.
- The benefits of methodological triangulation are discussed, including the validation of data, a more complete picture of discourse, and increased collaboration among researchers.
- The challenges of methodological triangulation are also acknowledged, including the time commitment, the need for methodological expertise, and the space limitations in publications.
- The implications of the findings for corpus researchers are presented, including the importance of methodological triangulation, the need to select appropriate methods and corpora, the importance of contextualizing research findings, the need for more replication-based research, and the need for more methodological commentary in corpus journals.
- The chapter concludes by emphasizing the importance of choosing the appropriate methods for corpus research and the potential benefits of methodological triangulation.

PS: to add to the mix

odological triangulation in corpus linguistics and its potential to improve the quality of research in this area.

At the beginning of this book, we asked whether all corpus methods lead to Rome. To continue the metaphor, we would conclude that most of them will lead to different parts of Rome. In other words, many of the findings detailed over the ten analysis chapters in this book were not shared, but were complementary—providing different parts of an overall jigsaw puzzle which enabled a holistic view of the nature of the Q+A corpus and its subsets. There were a few cases of shared findings, but they were relatively less frequent than the complementary ones and tended to involve only between two and four of the ten analysts. We thus note that, on the whole, different

References

- Baker, P. and Egbert, J. (eds.). (2016). Triangulating Methodological Approaches in Corpus-Linguistic Research. London: Routledge.
- Curry, N., Baker, P. and Brookes, G. (2023). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. Applied Corpus Linguistics, 4(1), 100082.
- Denzin, N. K. (1970). The Research Act in Sociology. Chicago: Aldine.
- Egbert, J. and Baker, P. (eds.). (2020). Using Corpus Methods to Triangulate Linguistic Analysis. London: Routledge.
- Gillings, M. and A. Hardie. (2022). The interpretation of topic models for scholarly analysis: An evaluation and critique of current practice. Digital Scholarship in the Humanities. doi:10.1093/llc/fqac075
- Gillings, M. and Mautner, G. (2023). Concordancing for CADS: Practical challenges and theoretical implications. International Journal of Corpus Linguistics.
- Marchi, A. and Taylor, C. (2009). If on a winter's night two researchers… a challenge to assumptions of soundness of interpretation. Critical Approaches to Discourse Analysis Across Disciplines, 3(1): 1–20.