CASS

The ESRC Centre for
Corpus Approaches
to **Social Science**

CASS POLICY REPORT SERIES

Unlocking the Power of Language for a Better Future

Our Approach to Driving Social Change

Welcome to CASS

At the ESRC Centre for Corpus Approaches to Social Science (CASS), we are passionate about understanding language and the impact it has on our world. As a globally recognised research centre, we use corpus linguistics to address real-world challenges. From healthcare and climate change to online safeguarding and education, our research helps shape policy, inform public debate and drive meaningful change.

We believe collaboration is key. Whether you are a researcher, policymaker, or organisation, we would love to work with you. Through our online and in-person training, we provide the tools and knowledge to help you incorporate corpus linguistics into your own work, whether you are just starting out or seeking to deepen your expertise. We also offer consultancy services, working with individuals and organisations to apply corpus techniques to key projects and policy development.

Vaclav Brezina
CASS Co-Director
Professor in Corpus Linguistics

Elena Semino
CASS Co-Director
Distinguished Professor

We hope this publication gives you a glimpse into the work we do and the ways we can

Leading the Way in Corpus Linguistics

Our Centre, based at Lancaster University, is a world leader in corpus linguistics: a powerful method that combines rigorous, systematic linguistic analysis with cutting-edge computational technology and specialised software to analyse vast amounts of language data.

Put simply, we study millions and billions of words from different areas of society – news articles, spoken language, social media, parliamentary debates and more – to identify patterns in how people communicate and interact with key issues. This helps us understand how language shapes society and how society shapes language.

With these insights, we advise policymakers, support organisations and provide evidence-based recommendations to improve public communication, shape policies and drive meaningful social change. In short, by identifying meaningful patterns in language, we help address some of society's most pressing challenges.

Since the 1970s, Lancaster has been a leader in corpus linguistics, conducting groundbreaking research that has shaped societal discussions. In recognition of this impact, the University received the 2015 Queen's Anniversary Prize for Higher Education. Our work spans areas such as healthcare, education, climate, hate speech, language testing and online security, collaborating with global partners to address real-world challenges.

Corpus Linguistics for Social Impact

The CASS team, led by Professor Vaclav Brezina, is at the forefront of pioneering research in corpus linguistics, developing innovative resources that are transforming the field. At the heart of this work are two major corpora – the British National Corpus 2014 (BNC2014) and the Hansard Corpus, alongside #LancsBox, an advanced software tool that has revolutionised language analysis.

These state-of-the-art resources provide researchers with unparalleled insights into how language evolves, how it reflects society, and how it shapes public discourse. Through large-scale linguistic data and advanced computational methods, CASS is making significant contributions to social science research, influencing policy, education and media analysis on a global scale.

A Window into Everyday Language

The British National Corpus 2014 (BNC2014) is a comprehensive dataset that captures current British English, making it a crucial resource for understanding contemporary language use.

BNC2014 has two key components. The Spoken BNC2014 captures real-life conversations across the UK, providing insights into regional dialects, evolving speech patterns and everyday communication styles. The Written BNC2014 includes a diverse range of written texts, such as books, newspapers, social media posts, blogs and SMS messages, showing, for instance, how written language is reflecting digital communication.

By representing diverse voices and communication styles, the BNC2014 allows researchers to explore linguistic inclusivity, generational language shifts and the impact of technology on communication. It has been instrumental in tracking how words, expressions and grammar evolve over time, helping educators integrate real-world language use into curricula.

Analysing Political Discourse Over Time

The Hansard Corpus, curated by CASS, is a unique linguistic dataset documenting over 200 years of UK parliamentary speeches, making it an unparalleled resource for studying political language and public debate. With over two billion words, it provides insights into how politicians frame key issues, how political rhetoric evolves and how language influences policymaking.

This corpus is essential for understanding how political discourse has shifted over time. Researchers can track long-term trends in rhetoric, policy discussions and public concerns, from the early 1800s to today. It is a valuable tool for policy analysis, showing how governments have addressed key issues such as healthcare, climate change and immigration. Additionally, it highlights power and persuasion, demonstrating how language is used to shape public opinion and influence decision-making.

Revolutionising Linguistic Analysis

#LancsBox is an innovative software tool for advanced linguistic analysis, used by over 250,000 researchers worldwide to process and explore large-scale linguistic datasets with unparalleled precision.

Its collocation analysis identifies frequently co-occurring words, revealing hidden relationships in language. Concordance searching allows researchers to analyse word usage across contexts, uncovering subtle shifts in meaning. Statistical and visualisation tools generate tables and interactive graphs, making complex language patterns easier to track and interpret.

Download #LancsBox for free to search the BNC20104, Hansard and other corpora or analyse your own data.

lancsbox.lancs.ac.uk

A Call to Action for Policymakers

The true value of language research lies in its ability to shed light on pressing social and economic challenges. To harness this potential, policymakers must recognise that investing in data and research infrastructure leads to more informed, impactful decisions that address real-world issues.

By fostering interdisciplinary collaboration, supporting ethical data practices and equipping researchers with the right tools, we can ensure that language analysis continues to drive meaningful insights, strengthening governance, shaping public policy and promoting social cohesion.

References

- Brezina, V. (2025). Corpus linguistics and AI: #LancsBox X in the context of emerging technologies. International Journal of Language Studies, 19(2), 75-90.
- Brezina, V. & Platt, W. (2025) #LancsBox X [software], Lancaster University, https://lancsbox.lancs.ac.uk
- Brezina, V., Hawtin, A., & McEnery, T. (2021). The written British National Corpus 2014–design and comparability. Text & Talk, 41(5-6), 595-615.

Optimising Vaccine Communication for Public Trust

Public health relies not only on medical science but also on how we communicate it. The success of vaccination programmes depends on more than just facts – it requires trust, understanding and effective messaging. Language plays a crucial role in shaping attitudes toward vaccines, influencing whether people choose to protect themselves and others.

Investigating Vaccine Messaging

Recognising this, the CASS team, in collaboration with the UK Health Security Agency (UKHSA), explored how vaccine messaging can be optimised to address hesitancy and promote public health. Led by Professor Elena Semino as part of the QuoVadis project, our research examined the role of metaphors – such as fire drill, raincoat and castle – to illustrate how vaccines work and why they matter.

What Works in Vaccine Communication?

Our study confirmed that brief health messages can positively influence vaccine attitudes. While metaphors did not significantly enhance or weaken this effect compared to literal language, they encouraged social discussions about vaccines, an essential factor in shifting public perceptions.

Crafting Effective Vaccine Messaging

Effective vaccine communication requires clear, engaging, and targeted messaging that adapts to different vaccine attitudes. Research suggests that emphasising personal benefits (e.g. self-protection) may resonate more with vaccine-hesitant individuals than messages focused on collective benefits (e.g. herd immunity). Personal stories, particularly those related to illness and protection, can be more persuasive than purely factual statements. Metaphors also help make vaccination discussions more accessible and relatable, supporting more effective public health communication.

Watch our short animations on vaccine metaphors on the SciAni YouTube channel.

youtube.com/@SciAni
(search "vaccine metaphors" in the channel)

A Call to Action for Policymakers

Vaccine education is not a one-size-fits-all endeavour. A single online message cannot overcome deep-seated hesitancy, but our research offers a cost-effective and scalable way to refine public health communication strategies.

To improve vaccine uptake, policymakers must invest in storytelling-driven campaigns, tailor messaging to specific audiences and continuously monitor public engagement with vaccine discussions. By combining scientific evidence with effective communication we can build a healthier, more resilient society.

References

- Demjén, Z., Brezina, V., Coltman-Patel, T., Dance, W., Gleave, R., Hardaker, C., & Semino, E. (2025). 'I am still unsure…'–Spontaneous expressions of vaccine indecision on Mumsnet. Applied Corpus Linguistics, 1-10.
- Flusberg, S. J., Mackey, A., & Semino, E. (2024). Seatbelts and raincoats, or banks and castles: Investigating the impact of vaccine metaphors. Plos one, 19(1).
- Semino, E., Coltman-Patel, T., Dance, W., Demjén, Z., Gleave, R., & Mackey, A. (2025). 'It's a shot, not a vaccine like MMR': A new type of vaccine-specific scepticism on twitter/X during the COVID-19 pandemic. Vaccine: X, 100620.
- Semino, E., Coltman-Patel, T., Dance, W., Deignan, A., Demjén, Z., Hardaker, C., & Mackey, A. (2024). Narratives, information and manifestations of resistance to persuasion in online discussions of HPV vaccination. Health Communication, 39(10), 2123-2134.

Transforming Language Education with Real-World Data

Language education is changing, with data-driven learning becoming an essential tool for improving how people learn and use languages. Instead of relying solely on textbooks, students and teachers now have access to real examples of how language is used in everyday life.

Corpus research, studying large collections of written and spoken language, has been at the heart of this transformation, providing valuable resources for learners and educators in the UK and beyond. At the core of this approach is corpus data, which serves as raw material in a language lab where students can explore real conversations, newspaper and academic articles, blogposts and other texts – just as scientists conduct experiments to understand the world.

Research led by Dr Dana Gablasova has shown that learning from real-world examples, rather than just focusing on rules and list of words, helps students communicate more confidently and effectively in practical situations. It also strengthens their critical thinking and analytical skills, making them more successful language users.

Empowering Students and Teachers with Corpus Resources

Our work has benefited a wide range of learners, from English as a Foreign Language (EFL) students to UK secondary school pupils studying A-level English Language. The Corpus for Schools project has provided teachers and students with access to British National Corpus 2014 (BNC2014) resources through BNClab, an innovative online platform that visualises corpus data. These resources have already reached hundreds of secondary school teachers and TESOL educators both in the UK and internationally.

Supporting English-Medium Instruction in Higher Education

The growing use of English as the main language of education, known as English-Medium Instruction (EMI), highlights the need for strong language skills to help students succeed at university. Our corpus-based research into EMI education addresses gaps in understanding students' real-life language use in these contexts, offering crucial insights to support students' academic performance.

Explore our Corpus for Schools project and access language learning resources.

lancaster.ac.uk/corpusforschools

A Call to Action for Policymakers

Policymakers must recognise the value of large-scale empirical approaches, such as corpus linguistics, in advancing language education and assessment. Integrating data-driven learning tools and evidence-based research into national education frameworks will ensure language education meets the needs of diverse learners.

By investing in data-driven language education policymakers can enhance language proficiency, academic success and inclusivity while creating greater opportunities for students across all backgrounds. A commitment to evidence-based policies will help ensure educational equity and better outcomes for all.

References

- Brezina, V., & Gablasova, D. (2024). A frequency dictionary of British English: core vocabulary and exercises for learners. Taylor & Francis.
- Gablasova, D., Harding, L., Bottini, R., Brezina, V., Ren, H. S., Iamartino, G., ... & Zottola, A. (2024). Building a corpus of student academic writing in EMI contexts: Challenges in corpus design and data collection across international higher education settings. Research Methods in Applied Linguistics, 3(3), 100140.

Key Terms Explained

Corpus Linguistics

Corpus linguistics is the study of language through the systematic analysis of large collections of real-world texts, known as corpora.

Using specialised software and computational methods, researchers identify patterns, trends and structures in how language is used across different contexts, such as news articles, social media, spoken interactions and historical texts.

By analysing millions or even billions of words, corpus linguistics reveals important linguistic and social patterns that would otherwise remain hidden.

Corpus

A corpus (plural: corpora) is a large, structured collection of real-world texts used for linguistic analysis. These texts can come from books, newspapers, social media, interviews, medical records, legal documents, or even spoken conversations.

Corpora can be:

- **General** (e.g. the British National Corpus 2014), reflecting broad language use.
- **Specialised** (e.g. the Hansard Corpus), focusing on specific domains like parliamentary speeches.
- **Synchronic**, capturing language use at a specific point in time.
- **Diachronic**, tracking how language changes over time through trend studies.

Concordance

A concordance is a key tool in corpus linguistics that lists occurrences of a word or phrase within a corpus, showing each in its immediate context. This helps linguists identify patterns, meanings and variations in language use.

Typically displayed in Key Word in Context (KWIC) format, the search term appears at the centre of each line, surrounded by words that provide context. For example, a concordance search for 'climate' in a news corpus might show the following:

The Paris	**climate**	accord is supposed to deliver precisely this type of global transformation
Because the impacts of	**climate**	change are already being felt in these countries
It's an important turning point in the Earth's	**climate**	

By analysing thousands of concordances, researchers uncover trends in meaning, collocations and discourse. Concordances can also be sorted, filtered and analysed statistically, making them valuable for both qualitative and quantitative research.

Keywords

Keywords are words that appear considerably more often in one corpus compared to another, making them stand out as linguistically and socially meaningful. Keyword analysis helps identify dominant themes, concerns or biases within a particular dataset.

For example, in a corpus of healthcare debates related to vaccination, words such as "MMR", "reaction" or "undecided" might appear considerably more often than in general language use, highlighting key areas of concern.

Collocation

Collocation is a systematic co-occurrence of words in corpora. These combinations of words show typical preferences in language. For example, the word 'people' commonly collocates with:

- **young** (more often than 'old')
- **other** (highlighting divisions between 'us' and 'them')
- **life** (capturing a broader association with everyday existence and human diversity)

Collocation is a linguistic and social phenomenon. It can be measured statistically, revealing meaningful linguistic and social relationships.

Frequency

Frequency refers to how often words or phrases appear in different contexts, measured using specialised corpus software. The repeated occurrence of specific words in particular contexts can shape public discourse and perception.

For example, corpus research on British newspapers has shown that discussions about people from different countries are often framed through the lens of immigration, frequently using terms such as "illegal immigration".

This repeated framing influences social discourse by drawing attention to a single issue while potentially overshadowing other important aspects.

Context

Context offers insight into the meaning of words, both in terms of their lexical sense and broader social implications. This is why context is crucial in corpus linguistics for analytical purposes (see Concordance).

Through the systematic analysis of the contexts in which words and phrases appear, corpus linguistics provides a unique perspective on how language reflects and shapes society.

For instance, corpus data reveal that the term 'elderly,' which dictionaries define as a polite way of saying 'old,' is frequently used in negative contexts, highlighting frailty, dependency and other vulnerabilities.

Annotation

Annotation refers to additional information added to corpora, allowing for more detailed and meaningful searches. This information can include:

- Grammatical categories (e.g. noun, verb, passive construction)
- Semantic categories (e.g. body parts, emotions, work and employment)
- Social categories (e.g. age, gender, socio-economic background)

By tagging texts with these categories, annotation enhances linguistic analysis, enabling researchers to explore language patterns with greater depth.

Best Practices for Curating and Analysing Social Data

Data is the backbone of corpus linguistics, and its ethical, sustainable, and rigorous management is essential for meaningful analysis, especially in the age of AI.

This checklist provides guidance for both data analysts and those who rely on linguistic and social data reports, including policymakers, journalists, and the general public, to help them understand the core principles of responsible and impactful social science research.

- **Ensure authenticity in data collection, including AI-generated data**

  Collect data that reflects real-world interactions, values, and preferences from diverse, representative sources. Where AI-generated data is used, critically evaluate it against human-produced data to maintain authenticity and societal relevance. Authenticity enhances the accuracy of insights into human behaviour and social trends.

- **Monitor data quality and ensure ethical use of technology**

  Implement rigorous quality controls to maintain data integrity, validity, and reliability. Use advanced computational technology transparently, with ethical considerations such as privacy, bias mitigation, and accountability guiding decision-making.

- **Focus on linguistic and social context**

  Consider both what is communicated and how it is expressed. Assess how linguistic choices and forms of expression shape social interactions and perceptions.

- **Emphasise frequency and language patterns in hypothesis testing**

  Use frequency of words, phrases, and contextual patterns as key elements in testing linguistic and social hypotheses.

- **Integrate AI responsibly with scientific and ethical oversight**

  Ensure AI applications adhere to core scientific principles (e.g. falsifiability), linguistic expertise (e.g. focus on context and frequency of use), and ethical guidelines (e.g. fairness, inclusivity, transparency, and accountability).

- **Foster interdisciplinary collaboration in data analysis**

  Encourage integration of insights from fields such as sociology, psychology, education, and economics to develop a more holistic understanding of social dynamics.

- **Make data insights accessible and impactful**

  Ensure data-driven findings are clear and accessible to policymakers, researchers, and the public. Foster open collaboration to maximise societal benefits and inform decision-making.

The ESRC Centre for Corpus Approaches to Social Science
Lancaster University
Lancaster, LA1 4YW
cass@lancaster.ac.uk
cass.lancs.ac.uk