UKRI Centre for Doctoral Training in
**Speech and Language Technologies**
and their Applications

# COM61003: Introduction to Responsible SLT Leadership ePortfolio

| | |
|---|---|
| Student Name | Aaron Fletcher |
| Student Registration | 230116573 |
| Cohort Number | 5 |
| Academic Year | 2023/24 |

**Contents**

**Ethical research practice** **3**
Epigeum Completion Certificate 4
**Responsible Research and Innovation (RRI) Guest Lectures** **6**
Lecture 1 7
Lecture 2 8
Lecture 3 9
Lecture 4 10
Lecture 5 10
Evidence of ORBIT Participation 12
Orbit PhD Technology Assessment report 13
**Personal Development Project (PDP)** **16**
**Entrepreneurship and Business Guest Lectures** **17**
Lecture 1 18
Lecture 2 19
Lecture 3 20
Lecture 4 21
Lecture 5 21
**Academic CDT / SpandH / NLP Seminar Series** **23**
Seminar 1 - NLP 24
Seminar 2 - NLP 25
Seminar 3 - NLP 26
Seminar 4 - NLP 27
Seminar 5 - NLP 28
Seminar 6 - NLP 29
Seminar 7 - Speech 30
Seminar 8 - Speech 32
Seminar 9 - Speech 33
UKRI CDT Conference Poster 35


UKRI Centre for Doctoral Training in
**Speech and Language Technologies**
and their Applications

# Ethical research practice

# Epigeum Completion Certificate

epigeum

## Certificate

Number: 1991230383

This is to certify that

**Aaron Fletcher**
of University of Sheffield

Successfully completed the course

**Research Integrity: Core**

| | |
|---|---|
| Good research conduct | 100% |
| Irresponsible research practices | 90% |
| Planning your research | 80% |
| Managing and recording your research | 80% |
| Data selection, analysis and presentation | 100% |

as part of the Epigeum Online Course System with a score of 88%.

Dated: 12 October 2023

epigeum

# Certificate

Number: 1991230383

This is to certify that

## Aaron Fletcher

of University of Sheffield

Successfully completed the course

## Research Integrity: Core

| | |
|---|---|
| Scholarly publication | 80% |
| Professional responsibilities | 90% |
| Communication, social responsibility and impact | 90% |

as part of the Epigeum Online Course System with a score of 88%.

Dated: 12 October 2023

# Responsible Research and Innovation (RRI) Guest Lectures

1. *Summary of each lecture attended (max 1 page). Include date, title, and presenter.*
2. *Evidence of participation and completion of ORBIT training*

*Add additional tables as required.*

# Lecture 1

| | |
|---|---|
| **Title** | Equality, Diversity and Inclusion (EDI) webinar – the role of EDI in enhancing research and innovation within a competitive engineering environment. |
| **Speaker** | Professor Jim McLaughlin |
| **Organisation** | Director of the Nanotechnology and Integrated Bioengineering Centre |
| **Date** | 29/9/2023 |
| **Narrative** | The presentation was a 1-hour online talk about equality, diversity, and inclusion. This was an external lecture (The lecturer was from the University of Belfast).<br><br>There were some short technical difficulties when starting the lecture, which resulted in poor lecture delivery, chiefly intermittent notification pings from the presenter's computer and slides not advancing correctly. This detracted from the presenter's ability to convey the lecture's important content. The online lecture format did not allow authentic audience interaction, so I could not ask questions.<br><br>The presenter started with a short introduction outlining what would be discussed, which I found to be a good way to determine quickly if this lecture suited the audience.<br><br>The lecturer used figures and relevant research well - each slide point had sufficient referencing. Notably, 90.7% of engineers were male, and 7.8% were BAME.<br><br>As part of a subgroup that EDI affects, more could have been done to highlight some of the issues these communities face. Real-world examples and case studies could have strengthened arguments. Additionally, I wanted to focus more on the lecture title (*'Enhancing Research and Innovation within a Competitive Engineering Environment'*) as little was discussed about what precisely a competitive environment is and how EDI will help improve that.<br><br>Ultimately, this lecture was designed to be a starting point for EDI within a research setting, with clear definitions provided, sometimes in excess, which hampered the development of more critical insights. It was great at outlining the key sectoral equality issues and the fact that if you measure EDI issues based on. who is working in your department, you don't necessarily understand the barriers which prevented people from working there in the first place.<br><br>This lecture's relevance to me will likely be in two main areas: ensuring my eventual PhD topic addresses and overcomes any real or potential EDI issues and my role in promoting research roles for other EDI-affected persons.<br><br>Take-home messages:<br>• In-person format preferred for giving lectures<br>• Overuse of definitions can detract from the message.<br>• Ensure technical issues are resolved before starting. |

| | |
|---|---|
| | • Use real-world examples to really strengthen arguments. |

## Lecture 2

| | |
|---|---|
| **Title** | Should a Robot Speak (if so, why, when and how)? |
| **Speaker** | Professor Roger Moore, Professor of Spoken Language Processing |
| **Organisation** | Department of Computer Science |
| **Date** | 17/10/2023 |
| **Narrative** | The presentation was a 1-hour in-person lecture on human-robot interaction by a world-renowned specialist.<br><br>The premise of the lecture was that researchers assume that interacting with a robot via speech is natural. Spoken language is the most complex form of communication we can see, and it has multiple components, not just automatic speech recognition systems, which need to be considered when creating these systems.<br><br>Roger presented an approach to understanding how language is perceived, such as pragmatics (the why did someone say this based on their environment, energetics (the motivational context, i.e. shouting) and Synhronics (such as directing attention or coordinating behaviours). I previously had started to consider that speech does not follow the traditional view (stimulus-response) and is much more complex; however, this lecture has formalised that approach - the coupled response.<br><br>He further explained that speech is more than just a single channel, as we utilise many questions to understand speech. He presented that each agent in a communication scenario has individual priors (i.e. a set of beliefs about the other agent in the communication), such as individuality, personality, demographics, and culture, which matter to the context. As ASR systems do not have those, and we as agents expect those things, this generates a mismatch of expectations and leads to confusion by the human agents. The question is, can ASR systems be created that approximate human agents so much that this mismatch of expectations can be bridged, or is it not possible?<br><br>Novel concepts - such as the habitability gap (a mismatch between capabilities and expectations of the human users) and likened it to the uncanny valley.<br><br>Take away points:<br>• Consideration of the end user is important when designing services<br>• Approximating humans might not be possible<br>• A holistic approach is needed when designing these systems. |

# Lecture 3

| | |
|---|---|
| **Title** | AI Technologies for Decision-Making: Challenges |
| **Speaker** | Prof Mirco Musolesi |
| **Organisation** | Department of Computer Science at University College London |
| **Date** | 18/10/2023 |
| **Narrative** | This hour-long online lecture was themed around the responsible use of AI technologies in decision-making scenarios, such as the economy, medical or other.<br><br>Decision-making systems are used in many social scenarios, such as conflict resolution and peace maintenance, with the potential risk of applying technology to this domain resulting in an impact on humans. One cited example is that the next financial crash is likely to emerge from AI.<br><br>The talk raised points such as these systems that don't have a vested interest in the outcome of these decisions (i.e., little ability to assess the human impact on the decisions they are generating).<br><br>The talk delved into assumptions made by AI creators, such as:<br>● Biological - the brain processes information using biologically equivalent on/off switches.<br>● Psychological assumption - that the brain can be viewed as a device operating on a known rule set.<br>● Epistemological assumption - All knowledge can be formalised.<br>● Ontological Assumption - the world consists of independent facts that can be formalised with precision.<br><br>The speaker outlined the decision-making process in machines and how it is one of the biggest problem areas, with generalised concepts such as machine autonomy, keeping humans in and out of the loop. Ultimately, decisions cannot be fair or equal unless the system has a stake in the outcome.<br><br>Relevance to me:<br>● Remember to look for unintended consequences of any code produced.<br>● Assessment of this risk needs to be continuous. |

# Lecture 4

| | |
|---|---|
| **Title** | Open Research Conversation: Sustainability and open data: Balancing environmental concerns |
| **Speaker(s)** | Tom Webb<br>Chris Olga |
| **Organisation** | Department of Bioscience at University of Sheffield<br>School of Advanced Studies University of London |
| **Date** | 21/2/2024 |
| **Narrative** | This lecture is part of Love Data week 2024.<br><br>Marine biology has a large amount of data available to it now, allowing it to answer more questions then possibly before. You can ask questions such as if the 2023 heatwave caused issued for marine animals. Tom has noticed that lack of data in marine biology is not a large concern as the data is readily available. They have developed a process for this, such as creating an occurrence record, which can be combined with other measures, such as temperature records for that area, so you can compare occurrence records, with temperatures to see if there is any correlation.<br><br>Different information on taxonomy and the characteristics of marine animals available. Loads of environment data available, typically through bio Oracle. Speaker talks about combining different sources of open source data to answer questions such as did the heat wave affect marine biodiversity.<br><br>Second speaker talked about their role from a humanities background to data science, stating that humanity subjects typically have no open datasets. Vast majority of literally scholars have to use data closed off by paywalls. Developed a toolkit that was developed, primarily aimed to remind people that digital work does have an impact on climate change. While IT is not the most significant problem with greenhouse gases, that was checked in 2021, and it likely to increase as AI developed. Transport contributes to 14% of green house gases, so need to consider when travelling to conferences etc. Toolkit is aimed at how to decarbonize aspects of their work, how to change systems within the university, carbon literacy etc.<br><br>Relevance to me:<br>• Need to consider the environmental impact when designing these sytems<br>• Should take steps to minimise technological debt. |

UKRI Centre for Doctoral Training in
**Speech and Language Technologies**
and their Applications

# Lecture 5

| | |
|---|---|
| **Title** | Social Media and Ethics |
| **Speaker(s)** | Nicolas Gold, UCL Computer Science |
| **Organisation** | UKRIO |
| **Date** | 21/2/2024 |
| **Narrative** | This 1 hour webinar lecture is part of UKRI online free webinar series, which looked at he the ethics issues for public data and providing researchers with an ethical defense for their principles and practice. The speaker was very clear that this was not a place to garner information regarding legalities or specific studies.<br><br>Anology of speaking a coffee shop was used, where conversations within a public space might be overheard. However, do the customers in the coffee shop consent to their conversations being listened to. People often have issues with this if their conversations are used for research into AI etc. For example, youtube users were suprised to be included in AI research (Hu 2019). It could be that the subjects, social media users, have a poor understnding of publically availabluty and implications so you need to consider the types of data, and risks presented to participants.<br><br>Informative statistics were presented by the speaker: On twitter, 61% of twitter users were not aware of their research use and 65% of people think that we shouldn't use tweets without permission. This shows a user desire for information about how their data is being used<br><br>The speaker raised the question of  if people trusted the social media platforms. Short answer no, and they do not come out well for trust in their data. A  study concluded users wanted meaningful transparency and we need to try to meet the expectations of the people's who data we use<br><br>Discuss on Public vs Private Data occurred, again with an analogy being given. In this foot path analogy, land owned by gatekeepers (such as employers for employers, Headteachers for school, Charities for service users, Social media companies for their users) and we as researchers have to ask permission from the landowners to user their data. He raised issues of expectation, consent, contract, regulation and access to data, and outlined how we should respect the stakeholders involved<br><br>Implied consent argument - If the service clearly explains the public disposition of that data and the available controls to a user, then the user gives implied constent when they post data.<br><br>Relevance to me:<br>● I will likely be engaged in some social media work during my PhD period and need to consider how to ethically best use this. |

UKRI Centre for Doctoral Training in
**Speech and Language Technologies**
and their Applications

# Evidence of ORBIT Participation

ORBIT

**Certificate of Attendance**

This acknowledges that

**Aaron Fletcher**

has successfully completed the course

**Foundations Course in Responsible Research and Innovation**

on the 11/06/2024

MARTIN DEHEAVER, DIRECTOR

UKRI Centre for Doctoral Training in
**Speech and Language Technologies**
and their Applications

# Orbit PhD Technology Assessment report

**RRI Maturity Report:** Technology-assisted review within the medical domain
**Name of Assessor**: Aaron Fletcher
**Date of Assessment**: 13/06/2024

**Use Case**
This PhD focuses on improving the time it takes to undertake a systematic review within the medical domain. A systematic review is a methodology for searching and appraising data for a given medical topic that is repeatable, reproducible, and considers all available evidence. Specifically, I will focus on information retrieval within the systematic review process and use datasets, such as the ones provided by Cochrane and Clef, to see if we can improve the recall of these methods by integrating large language models within the continuous active learning process. The use case for this evidence will be by clinicians who are undertaking these systematic reviews.

The Sustainable Development Goals Affected by this Project are...


Unquantified Positive Impact

The provision of good healthcare depends on synthesising available evidence into a format that encompasses all available data. If information is not retrieved that would have affected the outcome of a systematic review, this could lead to the provision of worse quality services.


Unquantified Positive Impact

Systematic Review information is used to teach a range of professionals, from students to updating seasons practitioners of medicine.

**Innovation Potential**
Based on your assessment of the SDGs affected by your technology and the following table, the Innovation Potential is 3

This is because your project impacted 2 Sustainable Development Goals. 0 of these were quantified, and 2 were unquantified.

| Number of SDGs affected by this technology. | Type of Impact | | |
|---|---|---|---|
| | No Impact | Unquantified | Quantified |
| 0 | Innovation Potential 1 | | |
| 1 | | Innovation Potential 2 | Innovation Potential 4 |
| 2 or more | | Innovation Potential 3 | Innovation Potential 5 |

**Technology Readiness Level**

The target technology readiness level for your project is :

- TRL 1 : Basic Research. Principles postulated and observed but no experimental proof available

**Calculation of RRI Intensity Level**

Based upon the Innovation Potential and the Target Technology Readiness Level for your project, the calculated RRI Intensity Level, based on the table below, for this project is:

- TRL 1 : Basic Research. Principles postulated and observed but no experimental proof is available

| | IP 5 | IP 4 | IP 3 | IP 2 | IP 1 |
|---|---|---|---|---|---|
| TRL 1 | RIL 5 | RIL 2 | RIL 2 | RIL 1 | RIL 1 |
| TRL 2 | RIL 5 | RIL 2 | RIL 2 | RIL 1 | RIL 1 |
| TRL 3 | RIL 5 | RIL 3 | RIL 3 | RIL 2 | RIL 2 |
| TRL 4 | RIL 5 | RIL 3 | RIL 3 | RIL 2 | RIL 2 |
| TRL 5 | RIL 5 | RIL 4 | RIL 4 | RIL 3 | RIL 2 |
| TRL 6 | RIL 5 | RIL 4 | RIL 4 | RIL 3 | RIL 2 |
| TRL 7 | RIL 5 | RIL 4 | RIL 4 | RIL 4 | RIL 2 |
| TRL 8 | RIL 5 | RIL 4 | RIL 4 | RIL 4 | RIL 2 |
| TRL 9 | RIL 5 | RIL 5 | RIL 5 | RIL 4 | RIL 2 |

Once the RRI Intensity level required for a particular project has been established by a research actor, be they prospective researcher or funding body, the question then comes as to what type of research project organisation has the right characteristics to deliver such an initiative. Clearly this evaluation covers a very wide range of parameters associated with research excellence etc, but we propose, that these parameters be extended to include due consideration of the RRI Maturity level of the organisation.

When combining these levels of RRI maturity with the categories and components of RRI, we arrive at a way of representing the RRI maturity model in the following form. We would like to underline that the differences between levels are not as clear-cut as the representation above may suggest.

**Calculation of RRI Intensity Level from SDG Impact and Target TRL**

Based upon the Innovation Potential and the Target Technology Readiness Level for your project, the calculated RRI Intensity Level, based on the table below, for this Project is 2

| | | |
|---|---|---|
| Level Five | Strategic | Organisation has adopted RRI as a component of its strategic framework and aims to ensure all R&D activities cover all (or most of) RRI components. |
| Level Four | Proactive | Organisation realises the benefits of RRI and seeks to proactively and increasingly integrate these into its business processes. |
| Level Three | Defined | Organisation has a definition of (components of) RRI and has integrated these into its business processes. |
| Level Two | Exploratory / Reactive | Organisation reacts to external pressures concerning aspects of RRI and experiments concerning appropriate processes. |
| Level One | Unaware | Organisation is not aware of RRI or its components and does not incorporate it into its processes. |

# Personal Development Project (PDP)

*Description of your PDP and plan for what needs to be done to realise the PDP project and how it will be carried out over the following 3 years. In addition to your ideas, target communities, goals, etc don't forget to include information about:*

- *Activities*
- *Timings*
- *Feasibility and risk*
- *RRI analysis*
- *Partners needed (if applicable)*
- *Funding needed (if applicable)*

*(3-5 pages)*

PDP proposal - Draft

# Entrepreneurship and Business Guest Lectures

*Summary of each lecture attended (max 1 page). Include date, title, and presenter.*
*Add additional tables as required.*

# Lecture 1

| | |
|---|---|
| **Title** | Early Career Research Through Commercialisation |
| **Speaker** | Various: Laura Talboters, Joe Carruth, Ryan Bramley |
| **Organisation** | University of Sheffield: Research, Partnerships & Innovation |
| **Date** | 23/11/2023 |
| **Narrative** | This was a 1 ½ hour online lecture with contrasting information to case studies.<br><br>The first lecture, given by Laura, was on how to apply research to commercial opportunities. She outlined previous schemes that have used CIP, which is no longer funded, and how it directly affects the current scheme (RISK) and why they believe it is better. She outlined what is available from the university, focusing on the available leadership skills/mentoring. Depending on the requirements, funding was available for ten projects, lasting 5 - 57 days. She focused on the impact of these projects and why they are a valid career choice for PGR / early researchers.<br><br>This led very well into the two case studies that were presented next. The first one, on contextualising captions for deaf users, was compellingly presented. They outlined their motivation, how the scheme (CIP) had helped them realise that and the tangible benefits that it had had for them (an ongoing researcher job with the university) and their target population (changes in how deaf subtitles were processed, and involvement of deaf people within their production, and evidence provided for improvement in legislation).<br><br>The following case study could have been presented better. No slide deck was available for this, and the result was an incoherent presentation that needed to show me why CIP had been integral to creating their business.<br><br>The best case study, given by the Phlux owner, was clear and concise. It showed their story of creating a material which had potential applications, how they raised funds to survey the market through a funding body, was given mentors and support from Sheffield University in this process and how they went on to develop a product which has market value, protects UOS IP and also can feed back into further research.<br><br>This set of lectures is an excellent reference for myself, and it has increased my knowledge and enthusiasm for commercial opportunities. |

# Lecture 2

| | |
|---|---|
| **Title** | Decoding value in business |
| **Speaker** | Speaker: Klaas Molapisi Chair: Rea Nkhumise |
| **Organisation** | PwC South Africa |
| **Date** | 24/1/2024 |
| **Narrative** | I attended an online virtual meeting that lasted for about 1 ½ hours on applying value to business. The focus of the meeting was on entrepreneurs, and the presenter and people within the group gave introductions. Instead of discussing the definition of value, the conversation revolved around what value means to society and how it relates to business.<br><br>The group talked about the principles or standards of behaviour that guide business and how business is a form of trade that should be focused on providing value. The discussion centred around the idea that values are what are applied to a business and how Apple provides value by making their products easy to use for creative purposes.<br><br>The group also talked about how working from a point of value can reduce corruption, grow ethics, and create a fair-trade market. There are three levels of values, ranging from corruption (the worst place to be), capitalism (more of a concern), to sustainability (the best place to be).<br><br>The main takeaway from the meeting was that businesses should prioritise working from a point of value rather than serving the investor at the cost of their values. When we have values, we establish clear principles, hold ourselves accountable, develop products from the point of impact, and create sustainable outcomes. Value doesn't necessarily create profitability, but it creates sustainability, ultimately what matters the most. |

# Lecture 3

| | |
|---|---|
| **Title** | AI Technologies for Decision-Making: Challenges |
| **Speaker** | Prof Mirco Musolesi |
| **Organisation** | Department of Computer Science at University College London |
| **Date** | 18/10/2023 |
| **Narrative** | This hour-long online lecture was themed around the responsible use of AI technologies in decision-making scenarios, such as the economy, medical or other.<br><br>Decision-making systems are used in many social scenarios, such as conflict resolution and peace maintenance, with the potential risk of applying technology to this domain resulting in an impact on humans. One cited example is that the next financial crash is likely to emerge from AI.<br><br>The talk raised points such as these systems that don't have a vested interest in the outcome of these decisions (i.e., little ability to assess the human impact on the decisions they are generating).<br><br>The talk delved into assumptions made by AI creators, such as:<br>● Biological - the brain processes information using biologically equivalent on/off switches.<br>● Psychological assumption - that the brain can be viewed as a device operating on a known rule set.<br>● Epistemological assumption - All knowledge can be formalised.<br>● Ontological Assumption - the world consists of independent facts that can be formalised with precision.<br><br>The speaker outlined the decision-making process in machines and how it is one of the biggest problem areas, with generalised concepts such as machine autonomy, keeping humans in and out of the loop. Ultimately, decisions cannot be fair or equal unless the system has a stake in the outcome.<br><br>Relevance to me:<br>● Remember to look for unintended consequences of any code produced.<br>● Assessment of this risk needs to be continuous. |

# Lecture 4

| | |
|---|---|
| **Title** | Open Research Conversation: Sustainability and open data: Balancing environmental concerns |
| **Speaker(s)** | Tom Webb<br>Chris Olga |
| **Organisation** | Department of Bioscience at University of Sheffield<br>School of Advanced Studies University of London |
| **Date** | 21/2/2024 |
| **Narrative** | This lecture is part of Love Data week 2024.<br><br>Marine biology has a large amount of data available to it now, allowing it to answer more questions then possibly before. You can ask questions such as if the 2023 heatwave caused issued for marine animals. Tom has noticed that lack of data in marine biology is not a large concern as the data is readily available. They have developed a process for this, such as creating an occurrence record, which can be combined with other measures, such as temperature records for that area, so you can compare occurrence records, with temperatures to see if there is any correlation.<br><br>Different information on taxonomy and the characteristics of marine animals available. Loads of environment data available, typically through bio Oracle. Speaker talks about combining different sources of open source data to answer questions such as did the heat wave affect marine biodiversity.<br><br>Second speaker talked about their role from a humanities background to data science, stating that humanity subjects typically have no open datasets. Vast majority of literally scholars have to use data closed off by paywalls. Developed a toolkit that was developed, primarily aimed to remind people that digital work does have an impact on climate change. While IT is not the most significant problem with greenhouse gases, that was checked in 2021, and it likely to increase as AI developed. Transport contributes to 14% of green house gases, so need to consider when travelling to conferences etc. Toolkit is aimed at how to decarbonize aspects of their work, how to change systems within the university, carbon literacy etc.<br><br>Relevance to me:<br>• Need to consider the environmental impact when designing these sytems<br>• Should take steps to minimise technological debt. |

UKRI Centre for Doctoral Training in
**Speech and Language Technologies**
and their Applications

# Lecture 5

| | |
|---|---|
| **Title** | Social Media and Ethics |
| **Speaker(s)** | Nicolas Gold, UCL Computer Science |
| **Organisation** | UKRIO |
| **Date** | 21/2/2024 |
| **Narrative** | This 1 hour webinar lecture is part of UKRI online free webinar series, which looked at he the ethics issues for public data and providing researchers with an ethical defense for their principles and practice. The speaker was very clear that this was not a place to garner information regarding legalities or specific studies.<br><br>Anology of speaking a coffee shop was used, where conversations within a public space might be overheard. However, do the customers in the coffee shop consent to their conversations being listened to. People often have issues with this if their conversations are used for research into AI etc. For example, youtube users were suprised to be included in AI research (Hu 2019). It could be that the subjects, social media users, have a poor understnding of publically availabluty and implications so you need to consider the types of data, and risks presented to participants.<br><br>Informative statistics were presented by the speaker: On twitter, 61% of twitter users were not aware of their research use and 65% of people think that we shouldn't use tweets without permission. This shows a user desire for information about how their data is being used<br><br>The speaker raised the question of  if people trusted the social media platforms. Short answer no, and they do not come out well for trust in their data. A  study concluded users wanted meaningful transparency and we need to try to meet the expectations of the people's who data we use<br><br>Discuss on Public vs Private Data occurred, again with an analogy being given. In this foot path analogy, land owned by gatekeepers (such as employers for employers, Headteachers for school, Charities for service users, Social media companies for their users) and we as researchers have to ask permission from the landowners to user their data. He raised issues of expectation, consent, contract, regulation and access to data, and outlined how we should respect the stakeholders involved<br><br>Implied consent argument - If the service clearly explains the public disposition of that data and the available controls to a user, then the user gives implied constent when they post data.<br><br>Relevance to me:<br>● I will likely be engaged in some social media work during my PhD period and need to consider how to ethically best use this. |

# Academic CDT / SpandH / NLP Seminar Series

1. *Summary of each seminar (max 1 page) including analysis of presentation quality (strengths & weaknesses). Include date, title, and presenter.*
2. *Details of when you presented your research topic as part of the main seminar series, CDT conference, or similar.*
   *Include a copy of the slides or poster used.*

*Add additional tables as required.*

# Seminar 1 - NLP

| | |
|---|---|
| **Title** | Language Resources and NLP Applications for Portuguese |
| **Speaker** | Professor Viviane Moreira |
| **Organisation** | Institute of Informatics at UFRGS |
| **Date** | 27/9/2023 |
| **Narrative** | This one-and-a-half-hour lecture covered two main topics: The professor's academic background and research areas (1 hour) and the explanation of the ARR (30 minutes).<br><br>The lecture started with a clear justification of why Portuguese presents a particular challenge for NLP analysis - chiefly that while Portuguese is the 6th largest language among native speakers, it only represents 2% of online web content. In comparison, English is spoken by 4% of the world and is represented by 55% of online web content.<br><br>The speaker briefly discussed their and their students' work, covering various topics. However, due to time constraints, there wasn't an in-depth discussion about any of the research. As a result, the presentation felt like a promotion of the department and students rather than a focused event. The abstract had suggested a more in-depth discussion, but this was not delivered.<br><br>The second part of the lecture, the ARR, was an overview of the submission and review process for various top-tier conferences. While I currently have no plans to submit, the information presented will undoubtedly become useful as I start to want to present at these conferences.<br><br>As NLP is my preferred field, it was interesting to get an overview of some current research topics; however, I gained little additional knowledge from this seminar about NLP.<br><br>The speaker was an L2 English speaker. However, they presented very clearly and concisely, struggling with enunciation with few words. The presentation style did, however, detract from conveying information, as the speaker's over-reliance on visual prompts distracted from communication.<br><br>Main takeaway points:<br>● Seminars with multiple students' research are inefficient for knowledge gathering.<br>● The submission process should be investigated closer to conference applications. |

UKRI Centre for Doctoral Training in
**Speech and Language Technologies**
and their Applications

# Seminar 2 - NLP

| | |
|---|---|
| **Title** | What deep nets can & can't do with language and why |
| **Speaker** | Prof. Robert Berwick |
| **Organisation** | MIT |
| **Date** | 4/10/2023 |
| **Narrative** | The presentation was a 1-hour talk in-person/online lecture about how deep neural nets currently understand language and why they can't perform like humans on this task. This was an external lecture (The lecturer was from MIT and was being held as part of the workshop on biologically inspired models of language).<br><br>The presentation style was one of the best I have seen (so far), particularly with the presenter's use of thorough examples and slide design stimulating active engagement. At no point did I feel that the lecture was repetitive and that the presenter added to the discussion about slide material rather than just repeating it. This lecture targeted post-graduates without defining basic terms and generally accepting that participants understood the domain (i.e., what a transformer is, and what a deep neural network is). This was appropriate, given the audience, which allowed a greater depth of material to be produced.<br><br>The first compelling argument presented was that deep neural networks cannot generalise if it is not within the training data, while children can. The example given was GoogleNet's inability to generalise from full-colour photographs to black and white drawings, while children can. As deep neural networks are essentially remembering a cloud of data, anything outside of that data cloud, such as black and white drawings, cannot be generalised to, and such, there exists two solutions: More training data to cover all these eventualities (in which case, is this general intelligence?) or that the neural networks are not the solution to human-equal intelligence.<br>One central argument by the presenter was that deep neural networks, such as chatGPT, do not universally approximate functions. This initially seemed bold, given that it is widely taught. The presenter presented some arguments for this case, such as rather than coordinate transformation of the space, it is divided into convex polytopes within the original space. If true, this currently is outside my means to understand the implications fully; however, it is an interesting point which I will look into further, and have been searching for related literature.<br><br>This is relevant to me, as, given the medical NLP domain that I intend to enter, there will be medical scenarios which have yet to occur before - new diseases being discovered or reclassification of diseases (which happens fairly frequently!).<br><br>This was a very informative lecture for me and has provided me with lots of references to consider:<br>● Logical syntax and semantics; their linguistic relevance https://doi.org/10.2307/410891<br>● Emergent linguistic structure in artificial neural networks trained by self-supervision https://doi.org/10.1073/pnas.1907367117<br><br>Take-home messages:<br>Understand and learn more about geometrical views of deep learning. |

# Seminar 3 - NLP

| | |
|---|---|
| **Title** | Evaluating Automated Citation Screening in Systematic Reviews: Metrics, Review Outcomes, and Datasets |
| **Speaker** | Wojciech Kusa, Final Year PhD candidate |
| **Organisation** | EU Horizon 2020 project DoSSIER |
| **Date** | 20./10/2023 |
| **Narrative** | The talker presented an hour-long talk on their work on systematic review automation in the medical domain.<br><br>His presentation used a mixture of discussion of their experiences and basic examples to outline their points. This proved very effective, especially with the complex topics he presented. He was an extremely fast talker, which meant that a lot of material could be covered, but I, as an audience member, could not comprehend his research findings.<br><br>His research area is something that particularly interests me, having previously been in the medical domain and conducted systematic reviews myself, yet even with this background, I found myself needing help to keep up with some of the concepts that he presented.<br><br>His argument on traditional evaluation metrics for systematic review paper generation (true negatives, work saved oversampling, confusion metrics, etc.) was particularly valuable. I wish he had spent more time on this topic. This felt like it could have occupied the entire time slot.<br><br>He then discussed the actual outcomes of systematic reviews and available datasets for this area, including advertising his soon-to-be-released dataset.<br><br>Overall, this lecture could have been less dense to communicate their ideas effectively. While I understand that the speaker was keen to share what they had learnt throughout their PhD experience, this felt like a series of talks on the topic area. The abstract did effectively communicate the topics to be discussed.<br><br>An interesting thing this presenter did was that, as the presentation was online, he aligned his webcam with his face, so it looked to the audience as if he was speaking to them. This felt effective.<br><br>Take-home messages:<br>● Consider reducing content to ensure the audience can follow along adequately<br>● Frequent checks on whether the audience follows through with examples or real-world scenarios.<br>● Enunciate and speak slowly, slower than my normal speaking pace. |

# Seminar 4 - NLP

| | |
|---|---|
| **Title** | Query Automation for Systematic Reviews. |
| **Speaker** | Harry Scells, Alexander von Humboldt Research Fellow. |
| **Organisation** | Leipzig University, Germany |
| **Date** | 03/11/2023 |
| **Narrative** | This presentation was a 1-hour remote presentation on the topic of query automation for systematic reviews. He was moderately successful in relaying the key concept: improvement within systematic reviews can come at different levels of the process.<br><br>It had some overlap with a previous NLP seminar which I had attended, which addressed how to evaluate the metrics of systematic reviews, so it felt somewhat complementary; however, this seminary was targeting systematic reviews at a higher level than before - through improving the search query of the initial screening and aiming to reduce the amount of time it takes for more in-depth reviews on papers. The contrast in approaches between the two seminars (automating research reviews vs reducing the number of papers to reviews) was great, as it showed how, from a single problem area, there are multiple different approaches which can be taken to enhance it.<br><br>During the lecture, the presenter used a simple image to explain the entire process of a systematic review. He emphasised the significance of automating this process due to its cost and time requirements. The lecturer also highlighted the importance of conducting a systematic review and explained in detail how it should be done. He outlined the different stages, which include screening through a boolean search, abstract screening, full study screening, study synthesis, and finally, the preparation and dissemination of systematic reviews. The process is aimed at refining studies from around 30 million to only about ten studies. However, it is an expensive process that takes up to two years to complete.<br><br>Harry's research centres around developing search strategies that can help reduce the need to screen large numbers of abstracts and full-text documents. Systematic reviews typically require complex queries that involve multiple subclauses a vast array of medical terms and ontologies that need to be carefully chosen. Harry's research aims to make the systematic review process more efficient and accurate by improving search strategy development.<br><br>Boolean queries are essential for systematic literature reviews, and they can be created automatically using either the conceptual method (which uses human expertise) or the objective method (which is more algorithmic). Both methods use seed studies for guidance and to balance the number of studies retrieved. The conceptual method involves gradually expanding high-level concepts, while the objective method follows a fixed order for each stage. Although the objective method is usually better, both methods can improve precision. |

UKRI Centre for Doctoral Training in
**Speech and Language Technologies**
and their Applications

# Seminar 5 - NLP

| | |
|---|---|
| **Title** | Taking stock of understanding and intelligence in LLMs... Then and now |
| **Speaker** | Allyson Ettinger |
| **Organisation** | Sheffield NLP group |
| **Date** | 26/1/2024 |
| **Narrative** | This was a 1 hour NLP seminar provided by the Sheffield NLP group. The lecturer started by examining existing approaches to evaluating LLMs and related work. The idea behind this was to provide a story of evaluation, which provided good structure to the presentation - past, present and future. She also discussed the historical pattern of pre-trained LLms of hype and bust, likely related to their beating evaluation tasks. However, upon closer inspection, they could improve on generalisation.<br><br>The lecturer outlined how current LLMs are performing exceptionally well on current evaluation metrics and how this is linked to their consumption of these metrics in their training data, which casts doubts on whether these evaluation metrics represent true learning (i.e. understanding of the context of a question, and applying it to different scenarios, or rote learning (i.e. memorisation of the relevant facts). Good, easy-to-understand examples were provided, with different tests showing property knowledge (Property knowledge and property inheritance).<br><br>She presented her research in the area, such as adding attractors (Items related to the question but not the answer) to problem tasks for pre-trained LLMs to solve, resulting in poorer performance. This reduction in performance did not match that of human performance reduction, suggesting that the pre-trained LLMs were extracting relevant information from the context of the question. She also explained that adding this distraction, particularly after the actual entity, further reduces performance.<br><br>A thought I had yet to consider, which was presented, was the lack of importance that negation provides with the subsequent word prediction, as there is a relatively high entropy of subsequent words that could be next. This is particularly important concerning natural language generation and could explain why hallucinations exist within these models.<br><br>The next section of the lecture focuses on what's next and how these evaluation metrics have changed. She presented solutions to the problem of pre-trained LLMs using contextual understanding to answer the question with the introduction of logic puzzles, which have been used in psychology previously.<br><br>Overall, the presentation style was excellent, if a little rushed, due to the amount of time allotted. I would be happy to receive further lectures from this person.<br><br>The major points that I took away from this lecture:<br>• Pre-trained LLMs can “cheat” evaluation tasks if they are trained on them<br>• Evaluation tasks are in an arms race against the models, we will likely require more complex evaluation tasks as time progresses.<br>• Just because something scores highly does not mean the model has good generalisability. |

# Seminar 6 - NLP

| | |
|---|---|
| **Title** | Demystifying prompts in Language models via Perplexity estimation |
| **Speaker** | Hila Gonen |
| **Organisation** | University of Washington |
| **Date** | 15/12/2024 |
| **Narrative** | This was an hour-long external lecture from a professor at the University of Washington, who has received two best paper awards in Corenll 2019. This lecture was provided via online meeting.<br><br>The topic for the seminar was very relevant and interesting to other projects I have undertaken recently - namely, the effect of prompting on large language models. Hila's presentation style was clear, particularly when exploring potentially challenging topics. It certainly catered to an audience with background knowledge of the topic area, as there was an assumption that perplexity was understood.<br><br>I would have improved this delivery by spending more time on perplexity and how it is derived. I had to search for information about perplexity while listening to the lecture, which made me need help understanding some of the later content.<br><br>Her topic and paper were highly interesting and novel, and it is one of the first talks I have been to that has attempted to quantify how prompt engineering works and even provide potential explanations as to why certain prompts work better than others. Using perplexity as a proxy, with the hypothesis that the lower perplexity of prompts results in more accurate scoring was great. I doubt its application to other models, as it does not necessarily match my observations - in that providing more data via the prompt (and hence increasing complexity) results in more accurate results. This could be a potential avenue for research; however, it is a challenge to work on the perplexity of a sentence when you don't have the corpora that the newer, closed models have been trained upon.<br><br>Hila's generation of different prompts is extremely relevant for the mini project, in which our subteam is looking at prompt engineering. Their approach is to generate 4-10 seed sentences, translate them into another language, then back to generate more, and finally ask a large language model to generate variations upon that, resulting in 100+ variations of the prompt. This did introduce noise, with some prompts not being grammatically correct; however, as her work showed, this doesn't necessarily impact the accuracy of the results.<br><br>Ultimately, this lecture was interesting and well-presented, and I particularly enjoyed her interaction with the viewers. In contrast, it was being presented, rather than awaiting for answers at the end of the presentation - it improved my understanding of the concepts! |

# Seminar 7 - Speech

| | |
|---|---|
| **Title** | Inclusive speech technology: Developing automatic speech recognition for everyone |
| **Speaker** | Dr Odette Scharenborg |
| **Organisation** | Associate Professor SpeechLab / Multimedia Computing Group |
| **Date** | 26/2/2024 |
| **Narrative** | The main focus of a one-hour lecture on speech technology was on inclusive speech technology and responsible research. Speech technology is primarily trained on everyday speech and normal voices, relatively homogeneous sources. This causes technology to be only available to a chosen few and increases digital divides, violating linguistic rights as included in the Human Declaration of Human Rights. The goal is to mitigate bias in speech recognition systems and uncover the causes of bias. There are multiple causes of bias, including differences in the training data distribution from the test data and algorithms and a need for more diverse development teams. The lecture discussed ways to reduce bias against diverse speaker groups in a SOTA ASR system, such as focusing on data scarcity, using data more effectively, and current research.<br><br>We have around 700 languages of the word; however, we only have 2% of that available as speech technology. Speech technology is primarily trained on normal speech and normal voices, a relatively homogeneous source.<br><br>There are many sources of diverse speech - such as regional accent, cold, and dysarthria, but speech technology is unavailable for these people. This is bad because technology is only for the chosen few; it increases digital divides and different access to information, especially for low/illiterate people, disabled people and people who speak 1 of the 98% of the world's language. Linguistic rights, as included in the Human Declaration of Human Rights, state that it is a human right to communicate in one's native language. Regardless of what it is, technology needs to include various languages. Ideally, we need lots of speech with many different transcriptions, so we need language models that predict the order of the words in the language. We need lots of text, which is unavailable in many languages. Transferring high-resource systems to low-resource systems often does not work. There is bias in data and systems against different speaker groups.<br><br>The goal is to uncover/mitigate bias in SOTA speech recognition systems. Imagine you have a test set of N=24 utterances from 4 speakers; 6/24 utterances need to be recognised. Imagine you have the same 25 % error rate; however, you are just misclassifying one type of language. This is biased against speaker groups. Depending on the database, sometimes female speech is better recognised than male speech. Speech from people between 18-30 is better recognised than speech from people outside. Adults > Children, Standard Accent> Regional accent, White speakers > Black Speakers. Different languages have different studies; however, there is a systematic difference between groups. |

| | |
|---|---|
| | There are multiple causes of this bias - firstly, if the training data distribution is different from the test data, you will have issues; algorithms tend to introduce bias, and you also need a diverse developers team.<br><br>How well can a SOTA ASE system deal with diversity of speech? First, you have to quantify it.<br><br>Quantification of the bias in SOTA ASR<br>What is the cause/origins of the bias?<br>Is bias dependent on the language?<br>Is bias dependent on the DNN architecture?<br><br>What is bias?<br>We do not know how to define bias; it depends on different factors; we define it as the difference in error rates between two speaker groups.<br><br>She looked at bias against gender, age, regional accents, and non-native accents.<br><br>Trained SOTA hybrid TDNNF-HMM and E2E asr (corpus gesproken nederlands) then tested on dutch jason corpus.<br>Found that female speech > male, teenagers > older adults > children<br>Native speakers > non-native speakers<br>Dutch regional speech > Flemish regional speech<br>Worse performance for all the adults - because regional accents may be more robust for adults.<br>Read speech > Human interaction speech.<br>Compared to the different architectures -> Differences in WER were smaller with hybrid systems.<br><br>How can we reduce bias against diverse speaker groups in a SOTA ASR system?<br>Focus on data scarcity<br>Using data more effectively.<br>Domain adversarial training<br><br>Data augmentation does not help with all improvements in speech ASR systems<br>More data with speed/volume perturbations<br>More speakers with pitch perturbations<br>More speakers with new articulation patterns will increase the variability of your training data.<br><br>To reduce WER and bias<br>-> Add natural or artificial data<br>-> Use data augmentation on standard speech.<br><br>If you use data augmentation and use data more effectively, you can reduce bias.<br><br>SOTA data augmentation and training techniques need to be generalised better for all types of diverse speech.<br><br>A low error rate does not mean you have a low bias. |

# Seminar 8 - Speech

| | |
|---|---|
| **Title** | Where Speech Technology Fits into AI Safety |
| **Speaker** | Jennifer Williams |
| **Organisation** | University of Southampton, South England Engagement Director, UK national EdgeAI Hub, Chair of ISCA security and privacy in speech communication |
| **Date** | 04/06/2024 |
| **Narrative** | This was a one-hour in-person lecture at the University of Edinburgh during the joint CDT UKRI conference. This was one of the few lectures that I have been able to attend in person, which was given in a lecture theatre in front of a sizeable audience. The presenter intended the talk to be non-technical, where they would not go into experiments and results, which I enjoyed less. I understood that the point was to be as general as possible, as the audience comes from an extensive background of knowledge/skill set; however, the audience was primarily students of CDT in NLP/Speech, so this was a missed opportunity.<br><br>The lecture started with the history of speech technology, which framed the subsequent lecture nicely, and the fast improvement in the capability of these models. It did, however, verge somewhat from the topic and used LLMs as an example of this increase - which, while they are related, and there have been some improvements with speech technology due to the transformer approach, I would have to look at this further to see the impact that this has had on the speech domain.<br><br>1. 2016 Merlin speech synthesis<br>2. Language grounding<br>3. Neural vocoders<br>4. Neural speaker representations (X-vectors) and style tokens<br>5. VoicePrivacy Initiative<br>a. Technical response to oncoming GDPR privacy concerns<br>6. Emotion recognition and paralinguistic analysis<br>7. Spoofing speaker verification<br>8. Partial spoofing detection<br>9. Multi-task neural networks (enhancement + recognition, speaker recognition, anti-spoofing)<br>10. Foundation models, ChatGPT, deep fakes and AI regulation<br><br>The presenters' point that speech technology faces more public scrutiny than ever before was well made. They mention that the increased public engagement with these technologies (even for deepfakes/novelty sake) means that there is a growing awareness of their capabilities. She referenced multiple challenges speech technology faces, such as opening up the public to novel threats and vulnerabilities.<br><br>A point that was particularly well made and resonant to my research area was that we tend to hold these LLMs/AI technologies to a higher standard than that we have of existing technologies, such as internet search. Both can return answers to questions we might not want to be freely available (such as how to build a bomb). However, internet search is almost expected to be able to return these illicit results, while AI models could be better if they do. This dual standard of accountability for these technologies is something that I want to explore further in future lectures within RRI. |

UKRI Centre for Doctoral Training in
**Speech and Language Technologies**
and their Applications

# Seminar 9 - Speech

| | |
|---|---|
| **Title** | AI analysis of Voice to Aid Laryngeal Cancer Diagnosis |
| **Speaker** | Mary Paterson |
| **Organisation** | University of Leeds |
| **Date** | 12/06/2024 |
| **Narrative** | This was a one hour lecture on AI analysis of laryngeal cancer. General symptoms are a change in voice, dysphagia etc - they are very broad symptoms. 160,265 cases in men since 2021, 24,350 in female.<br><br>The vast majority of these cancers are stage 1 / stage 4. Median survival time after stage 4 for 5 years is 30% verses 90% at stage one. Stage 4 usually necessitates removal of the voice box + long term radiation therapt. Lots of side effects of having this treatment. Ultimately we want this dignosed at stage 1 to reduce this harm.<br><br>2 week wait pathway. Start at PCP, 2 weeks wait then go to a specialist. The general idea is that using an AI system is inbetween the PCP + specialist, looking at classification / prioritisation of patients. SPecialists are seeing a huge number of patients but actually rate of having cancer is really low (5%). This can ease the load on specialists.<br>Pinpoint test system.<br><br>Applicaiton possibilities -> recording device in PCP, and recording at home. If using recording at PCP we have a more standardised system, which can be human vetted before. Barrier of entry to use a smartphone/having an environment that won't cause issues with classification/diagnosis.<br><br>Started with feature extraction - MFCC - eGeMAPSv02- wav2vec2 feature states. Wav2Vec2 feature states were not interpretable.Wav2Vec2 performed better than MFCC?eGeMAPSv02.<br><br>Potential contributing factors for the failure of the models on the external datasets.<br>- Patient demographics<br>- Recording length<br>- Recording devices<br>- Recording environments.<br><br>Did the length of the recording affect the model, so shortened the healthy patients?<br>● MFCC - Shorter recordings did worse, longer recordings did better (generally).<br>● eGeMAPS2-v02 o not a big difference<br>● Wav2vec2 - needs around 5 seconds to get good performance. However it did not translate to the external dataset.<br><br>Wanted to improve this:<br>● Investigate preprocessing to reduce the degradation on external data.<br>● Investigate more complex speech tasks.<br>● Investigate the effects of recording device.<br><br>Audio was collected from patients using the cookie theft test. Used to reduce the chance of speaker revealing identifiable data. Recorded on two devices and a Nokia 110 phone. Collected more information for demographics, such as Age, Gender, BMI, smoking status, Caffeine and Alcohol consumption.<br>Often get people who don't tick the boxes the for all symptoms. Because the data is patient reported, it is often is noisy.<br>Inbuilt noise reduction software on phones can reduce the amount of data collected - as it often treats elongated vowels as background noise after a certain point. Differences in the devices exist and is |

| | |
|---|---|
| | easily found.<br><br>The lecture was presented in a great and informative way, with high interactivity. uestions: Have the patients consented to patient being shared. |

# UKRI CDT Conference Poster

## TECHNOLOGY ASSISTED REVIEW
within medical domain

University of Sheffield

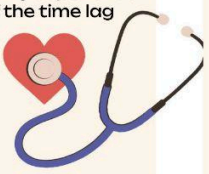
**Author**
Dr. Aaron Fletcher
ahafletcher1@sheffield.ac.uk
**Supervisor**
Dr. Mark Stevenson
mark.stevenson@sheffield.ac.uk
University of Sheffield

SLT Centre for Doctoral Training

UK Research and Innovation

### ABOUT

As a **veterinary surgeon** with a decade of medical practice, my background may seem unusual for pursuing a PhD in NLP. During my practice, I aimed to improve the medical evidence base through **publishing a PICO review** on non-steroidal anti-inflammatory drugs and their potential to mitigate gastrointestinal side effects. My experience in medical research and awareness of the time lag between **conducting reviews and disseminating findings** motivated me to work on facilitating clinicians' access to the latest evidence through NLP.

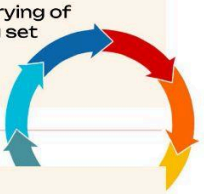


### KEY TERMS

**Systematic Reviews** aim to identify, evaluate and collate findings relevant studies to a specific health issue, to make the available evidence easier to use.

**TAR** is the iterative process of retrieving documents from a collection, and then reviewing them until a substantial majority, or threshold has been met.

**Active Learning** is Interactive querying of the collection, adding to the training set with new labeled data.



### AREA

The medical evidence base is expanding at an unprecedented rate, with a staggering **268.93% increase in publications from 2002 to 2022**. This growth poses a significant challenge, as it becomes impractical for a single human to thoroughly review all evidence returned from queries in a reasonable timeframe. Nevertheless, this is precisely what medical researchers are expected to do during the peer review process. As this body of knowledge continues to swell, the need for efficient tools that can review documents becomes increasingly urgent

To date, medical TAR research has focused on using techniques such as logistic regression with active learning. Could we **leverage more SOTA models,** such as Mamba or GPT, which have greater capacity for transfer learning, the ability to understand complex queries, improved text understanding, to achieve better medical TAR results? There is a plethora of AL techniques which can be used during the TAR process, however, **are some** more **superior** than others **within the medical dataset?** If so why?

### OBJECTIVES

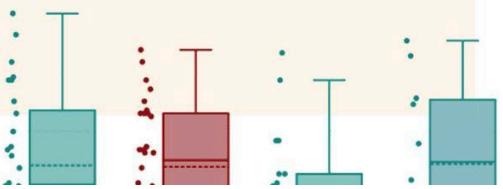
- Adapt SOTA models (e.g. Transformers/Mamba) for the Active Learning TAR approach within systematic reviews.
- Explore the effect of different active learning approaches on these models.
- Explore prompt engineering approaches on LLMs within TAR

### DATASETS

**CLEF eTAR** is a collection of diagnostic test accuracy and intervensional systematic reviews.

**Cochrane** is a large collection systematic reviews.

**PubMed** is the largest collection of medical evidence.



This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1]