Institution: University of Wolverhampton

Unit of Assessment: 26 Modern Languages and Linguistics

1. Unit context and structure, research and impact strategy

1.1 Unit Context and Structure

In REF 2021, the Modern Languages and Linguistics (UoA 26) submission from the University of Wolverhampton is based on the activities of its Research Group in Computational Linguistics (RGCL). RGCL is a coherent unit within the Research Institute in Information and Language Processing (RIILP), alongside the Statistical Cybermetrics Research Group (SCRG). The Head of RGCL is also the Director of RIILP. RGCL comprises 12 (9.7 FTE) research staff and eight PhD students. RGCL was entered in RAE2008 and REF2014. It has expanded in terms of FTE staff (9.7 vs. 7.8 in 2014), doctoral degrees awarded (10 vs. eight in 2014), research income raised (GBP4.3m vs. GBP3.6m in 2014), and peer-reviewed research outputs produced (314 papers vs. 119 in 2014).

1.2 Evidence of the Achievement of Strategic Aims for Research and Impact During the Assessment Period

RGCL set six strategic aims to increase the quality and breadth of the Group's research, to expand in terms of staff, develop new areas of expertise, and generate funding to support the Group's activities. Below we detail the achievements against each aim. The parentheticals "(ECR)" and "(PGR)" indicate Early-Career Researchers and Postgraduate Research Students, respectively. The terms "natural language processing" (NLP) and "computational linguistics" are used interchangeably.

1.2.1 Aim 1: Pursuing Computational Linguistics research in areas where research outputs will clearly benefit society

During the REF assessment period, RGCL pursued seven main lines of research:

- *NLP Methods and Tasks* [Mitkov; Evans; Ha; Orasan; Can];
- *Corpus-Based Studies* [Hanks; Oakes; Moze; Corpas Pastor; Mitkov; Evans; Orasan; Taslimipoor (ECR); Can];
- *NLP for Digital Humanities* [Mohamed; Oakes; Sarwar; Zampieri, Mitkov];
- *NLP for Technology-Enhanced Learning* [Ha; Mitkov; Yaneva (ECR); Can; Oakes; Sarwar; Evans];
- *Language Processing for Assistive Technologies* [Evans; Orasan; Yaneva (ECR); Can; Mitkov; Ha; Blain];
- *NLP for Social Media* [Zampieri; Taslimipoor (ECR); Orasan; Sarwar; Can; Oakes; Evans; Mitkov];
- *Translation Technology* [Mitkov; Orasan; Blain; Corpas Pastor; Taslimipoor (ECR); Zampieri; Ha; Mohamed].

RGCL's achievement of Aim 1 included the development of:

- new methods for anaphora resolution, named-entity recognition, shallow and partial syntactic parsing, morphosyntactic analysis, speculation and negation detection, and text normalisation; and development of deep learning approaches for representation learning (NLP Methods and Tasks);
- new lexicographic resources and NLP tools to assign meaning to verbs and verbal multiword expressions during its coordination of the AHRC-funded project "Disambiguation of Verbs by Collocation" (*DVC*, 2012-2015), as well as automatic detection and translation of multiword expressions and other corpus-based phraseology studies (Corpus-Based Studies);

- NLP methods for stylometry, author profiling, plagiarism detection, and movie subtitle content analysis; and development of the first Arabic Biographical Dictionary (60,000 entries) (NLP for Digital Humanities);
- NLP methods to improve assessment of educational attainment for medical licensure (NLP for Technology-Enhanced Learning);
- new user-focused assistive language technologies for people with autism spectrum disorder (ASD), including detection methods (Language Processing for Assistive Technologies) and their evaluation;
- next-generation NLP-based translation technology to assist professional translators; semantically-enhanced translation memories and quality estimation methods; and coordinating the European project "Exploiting Empirical Approaches to Translation" (*EXPERT*) (Translation Technology).

The activities described in Sections 1.4-1.7 are further indicators of the unit's success in achieving Aim 1. This led to RGCL's development of numerous open access research outputs, including 36 NLP tools, 19 datasets, and 314 peer-reviewed academic papers, a reflection of the open research environment fostered by RGCL. RGCL's research output during the census period covers 106 different topics, including 76 journal articles, seven books, 48 book chapters, 183 papers in peer-reviewed international conferences and specialised workshops, five edited volumes, and eight workshop proceedings.

1.2.2 Aim 2: Fostering collaboration with leading researchers in other institutions and with non-academic users, especially from companies

Section 4 details the methods used by the unit and cites examples of collaborations undertaken to achieve this aim in the REF period.

1.2.3 Aim 3: Setting up new Master's programmes in Natural Language Processing (NLP) research areas

RGCL achieved this aim and is successfully running three master's programmes: *Computational Linguistics*; *Practical Corpus Linguistics for ELT, Lexicography and Translation*; and an *Erasmus Mundus Master's in Technology for Translation and Interpreting* (EMTTI). All RGCL teaching is underpinned by research; RGCL's work in translation technology attracted funding to appoint specialist researchers Blain, Mohamed, and Zampieri. Their contributions and Mitkov's seminal work in the field have led to external funding of RGCL's EMTTI programme. Globally, EMTTI is the first Master's programme in Technology for Translation and Interpreting, which focuses on research to meet the practical needs of this industry. RGCL's success in establishing and participating in master's programmes contributes to its sustainability.

1.2.4 Aim 4: Maintaining the quality of the unit's research to the highest standards

The quality of RGCL's research is evidenced by its research outputs, successful project proposals, and invitations to collaborate on research and present keynote speeches.

During the REF period, RGCL maintained the quality of its research to the highest standards. This has been strengthened through RGCL's exposure to cutting-edge research via:

- the organisation and leading of four NLP-orientated seminar series: Natural Language Processing, Technologies for Translating and Interpreting, Machine Learning/Deep Learning, and Digital Humanities (https://bit.ly/38QNiid);
- more than 50 collaborations in academia and beyond contributing to 106 research topics;
- provision and use of a generous conference fund;
- organisation of 17 international conferences and workshops;
- running of three Master's programmes and supervision of 12 PhD students;
- provision of support for training and summer school attendance.

The four RGCL seminar series demonstrate the vitality of the research environment.
RGCL staff engaged with recent developments in the field through conference attendance, which also provided opportunities for peer interaction and collaboration. RGCL's success and involvement in NLP shared tasks (in which research centres from around the world compete to develop practical solutions to pressing NLP challenges) also demonstrates its achievement of Aim 4. RGCL was ranked first in international shared tasks at the Parseme-2018 and WMT-2020 conferences, with the Machine Translation quality estimation system developed by RGCL at the latter being superior to those developed by any of more than 50 research groups from all over the world. In 2019, RGCL organised a Summer School on Deep Learning in Natural Language Processing, including practical sessions delivered by Taslimipoor (ECR) and Rohanian (PGR). Organisation of, and participation in, summer schools and shared tasks maintain RGCL's engagement in cutting-edge research.

1.2.5 Aim 5: Generating increased external funding by identifying new sources of income

In the current REF cycle, RGCL successfully achieved Aim 5 by identifying new sources of income. It raised GBP4.3 million, an increase from the previous REF period (GBP3.6 million); see Section 3.1.

1.2.6 Aim 6: Maintaining the unit's cohort of PhD students and supervising them to successful completion

In this REF period, RGCL successfully maintained its cohort of PhD students and supervised them to successful completion. In 2014-2021, 12 PhD students were appointed. RGCL has an average annual PhD cohort of nine students, with 10 successful completions in this REF cycle compared to eight completions in the previous one; see Section 2.2.

1.3 Research Impact

During the current REF period, RGCL's impact strategy was based on:

i. focused applied research and development to address the needs of service providers for people with ASD, psychometricians, educational assessment experts, humanities scholars, translators, and other potential users;
ii. motivated scientific, industrial and social networking;
iii. dedicated consultative training.

This strategy underpinned RGCL's success in achieving impact in the current REF cycle. Collaboration between RGCL researchers and research beneficiaries brought new insights into end-user requirements, evidence-based derivation of new approaches to address user needs, and deployment of user-focused evaluation methods to ensure relevance of approaches (NBME ICS). User participation in an EC-funded project to develop a "Flexible Interactive Reading Support Tool" (*FIRST*) brought input from end users into the development and evaluation of the NLP tools, an example of co-creation of knowledge by partners and service users from several sectors (FIRST ICS).

1.4 Strategic Aims for Research and Impact in the Next REF Cycle

RGCL will focus on new interdisciplinary research topics with an emphasis on five specific research goals:

G.1 Develop innovative NLP methods and applications to benefit academia and wider society.
G.2 Further the application of NLP and the latest deep learning techniques in Digital Humanities. This will include applications in stylometry (including plagiarism detection, forensic linguistics, and native-language identification).
G.3 Explore the development, application, and evaluation of a wider range of deep learning approaches to address RGCL's current and future research topics. This will include the application of the latest advances in deep learning, transfer learning and multitask learning in applied NLP and Translation Technologies.

G.4 Pursue research in multimodal processing, including eye tracking and image processing to support evidence-based and user-focused development and evaluation of applied NLP.
G.5 Pursue NLP and Translation Technology research to support a wider range of languages, including under-resourced languages such as Sinhalese, Nepalese, Turkish, and Urdu.

RGCL will continue its *NLP for Digital Humanities* research by developing new methods to assess information relevance and adding new lines of research on author profiling, offensive language detection, and stylometry. The unit also aims to submit project proposals on Arabic literary analysis (Doha Institute for Graduate Studies) and ethics research (Research Center for Islamic Legislation and Ethics). In *NLP for Technology-Enhanced Learning*, RGCL will seek to develop new approaches for automatic test-item generation using synthetic data, distractor generation, short item scoring, and prediction of item characteristics, further enhanced using deep learning techniques. In *NLP for Social Media*, RGCL will collaborate with SCRG for social-web data collection and sentiment analysis. It will also aim to develop improved methods for hate speech and fake news detection and mitigation of bias in social-media derived language models. RGCL will continue its research in *NLP Methods and Tasks* and *Translation Technology*, pursuing topics in which it has already produced internationally-leading research (e.g. anaphora resolution). Future research topics will include application of deep learning methods to develop new-generation translation technologies and quality estimation methods, and applied research on multimodal processing (e.g. eye tracking) for user-focused evaluation.

In 2021-2026, we will expand our strategy for research impact by focusing research outputs on selected areas likely to benefit society. In most NLP research areas, impact depends on identifying user requirements, developing solutions in consultation with end users, and involving users in the evaluation of those solutions.
RGCL will seek to identify larger numbers of potential beneficiaries of its research by:

- conducting research addressing limitations of current evaluation methods, which often lack relevance to end users;
- exploiting research on multimodal processing to accurately identify the most impactful applications of NLP research;
- providing open access to our NLP tools and datasets and attracting potential beneficiaries through dissemination via social media and relevant mailing lists;
- disseminating information about RGCL's applied NLP research among its growing research networks, including that of the EMTTI Master's programme which now has more than 25 industrial associated partners. This will contribute to the training of next-generation translators and interpreters and identify the future needs of the translation and interpreting industry;
- organising and participating in shared tasks to raise awareness of the availability and efficacy of RGCL's research outputs (datasets, NLP tools).

1.5 Support for Interdisciplinary Research

Computational Linguistics/NLP is inherently interdisciplinary, involving collaboration with computer scientists, linguists, mathematicians, psychologists, and translators. Furthermore, one of RGCL's research lines is *Applied NLP*, which involves end users in project preparation, development, and evaluation. In *FIRST* (Mitkov, Orasan, Evans, Yaneva), RGCL collaborated with clinicians, psychiatrists, computer scientists and linguists. The unit's (Ha, Mitkov, Yaneva, Taslimipoor, Evans) collaboration with the USA's National Board of Medical Examiners (NBME) involves psychometricians and medical practitioners, whilst one of RGCL's primary research lines is based on collaboration with translators and interpreters (Mitkov, Orasan, Corpas Pastor, Blain). RGCL collaborated with the University of Quebec, applying NLP in sociological research (Mohamed) and with the Austrian Centre for Digital Humanities in a project applying NLP methods for prosopographical research (Mitkov, Orasan, Zampieri, Evans). Most of RGCL's proposals for external research funding are explicitly interdisciplinary. RGCL provided funding for eye-tracking equipment and conference attendance to support award-winning interdisciplinary research on psycholinguistics and accessibility (Yaneva).

RGCL frequently hosts (up to 12 months' duration) guest researchers, including professors and ECRs. Visitors have included linguists, mathematicians, statisticians, computer scientists, educators, psychologists, health-service providers, translators, and interpreters. These have collaborated with RGCL staff on research projects, given presentations and guest lectures, and co-supervised PhD students.

1.6 Progress Towards Becoming an Open Research Environment

RGCL's progress towards creating an open research environment is demonstrated by the number of open-access language resources, NLP tools, and peer-reviewed research papers produced in this REF period (Table 1).

Research line	Selected datasets	Selected NLP tools	Peer-reviewed research papers
NLP Methods and Tasks	1 (Token sequences labelled with shallow syntactic information)	1 (Identification of verbal multiword expressions)	65
Corpus-Based Studies	7 (The Pattern Dictionary of English Verbs (PDEV); four comparable and two parallel corpora)	4 (Corpus analysis and processing)	32
NLP for Digital Humanities	1 (Linked idioms in five languages)	4 (Segmentation and morphosyntactic analysis of Arabic)	58
NLP for Technology-enhanced Learning	-	2 (Multiple-choice question distractor generation; feature extraction indicating response times in educational assessment items)	19
Language Processing for Assistive Technologies	1 (Autistic eye-tracking corpus)	1 (Sentence simplification)	30
NLP for Social Media	1 (Offensive microblogs corpus)	2 (Detection and categorisation of irony; offensive language detection)	26
Translation Technology	2 (Corpus annotated for translation memory cleaning; editing operations of four professional translators)	3 (RNN-based translation evaluation; paraphrase-based translation memory segment retrieval; semantically-enhanced translation quality estimation)	95
All	13	17	325

Table 1: Open-access research outputs

Software and language resources (corpora; dictionaries) developed by RGCL are available on the Group's website (http://rgcl.wlv.ac.uk/resources-2/), via dedicated portals (e.g. http://pdev.org.uk/), and in public repositories such as GitHub. TransQuest, our Machine Translation quality estimation tool, was downloaded 531 times in July 2020 alone, after being ranked No. 1 at WMT-2020 the same month. In the REF period, the unit provided open access to 19 datasets and 36 NLP tools. As well as enhancing the vitality of the unit, these also facilitate research and collaboration, thus contributing to sustainability.

1.7 Support for a Culture of Research Integrity

RGCL offers continuous training in research integrity in computational linguistics to both staff and students. This begins in their first year of appointment. RGCL members are regularly updated about changes to relevant policies. Research integrity goals are included in staff appraisals, meetings between research students and supervisors, and the annual progress reports of PhD students.

RGCL's open research environment helps to ensure the transparency, verifiability and replicability of its research, contributing to the culture of research integrity.

2. People

RGCL's staffing has increased from 9 (7.8 FTE) staff in REF 2014 to 12 (9.7 FTE), contributing to the sustainability of the unit and enhancing its research vitality. RGCL's research, teaching, and administrative activities (including fundraising) are implemented by research-active academics (three professors, of whom two are part-time, two Readers, two senior Lecturers, and five Lecturers) and five administrators. More than 90% of RGCL's research-active staff are on permanent contracts, as are 80% of its administrators. RGCL currently supervises 10 PhD students.

The Group aligns itself with the University Concordat to Support the Career Development of Researchers by encouraging all staff to seek personal and professional development and recruiting strategically to strengthen the Group. We encourage interconnected research between all academic staff, thus creating a supportive and inclusive culture. Annual progress reviews are used as a springboard to embed career development through mentoring, providing support for training at all levels and accommodating flexible working requests wherever possible. All staff and students benefit from their participation in reading groups which provide the opportunity for collaborative creative thinking sessions across RGCL's research areas.

2.1 Staffing Strategy and Staff Development

RGCL's approach to staffing and staff development is geared towards the attainment of its strategic aims and goals for research and impact in 2021-2026 (Section 1.3).

Staffing Strategy

RGCL's staffing strategy includes nurturing and supporting ECRs: several PhD students have been employed as Research Associates/Fellows after their successful completion (e.g. Yaneva, 2016; Taslimipoor, 2018). Contributing to the application of NLP to technology-enhanced learning, ECRs Taslimipoor and Yaneva were initially appointed as research associates with expertise in deep learning, phraseology, and distributional semantics (Taslimipoor), and psycholinguistics and text accessibility (Yaneva).

The unit encourages and strengthens teaching by promoting promising ECRs to Lecturers with significant responsibility for research (Moze in 2018; Yaneva in 2019; Evans in 2020). Zampieri was appointed as a Research Fellow in 2017 and Mohamed as a Senior Lecturer in 2018. They contributed their collective expertise in language and dialect variation, second-language acquisition, educational NLP applications, author profiling, translation technology and multilingual text analysis for under-resourced languages. These topics had been identified as being of strategic importance in RGCL's plan for research and impact in 2014-2021.

When recruiting new staff, candidates are assessed in terms of their track record in research, research potential, areas of expertise, and personal research interests. As a result, RGCL has successfully recruited outstanding academics who are capable of independently producing world-class research whilst enhancing the unit's established research areas.

In the past 12 months, RGCL appointed a Reader (Can) and two research-active Lecturers (Blain and Sarwar), whose expertise will contribute directly to RGCL's strategic aims and goals for research and impact in deep learning, translation technology (Blain); deep learning, morphosyntactic analysis, named-entity recognition and sentiment analysis (Can); and stylometry, authorship attribution and native-language attribution (Sarwar).

Between 2013 and 2020, RGCL benefited from internal University funding (*RIF1* and *RIF3*), which was used to expand the group through the appointment of a Reader, a Senior Lecturer, and two Research Fellows in Computational Linguistics and Translation Technology.

Staff Development

Staff development is central to RGCL's research strategy and sustainability. RGCL staff are supported from their first appointment as ECRs to their later progression as research specialists and their continued development through senior academic levels. All RGCL staff benefit from a range of structures and incentives to support research. This includes staff development programmes, the Early Researcher Award Scheme (ERAS) which provides grants to support ECRs, and mentoring. In this REF cycle, three of RGCL's ECRs obtained ERAS funding. Research staff at the mid-point or later in their careers are encouraged to apply for targeted university development programmes. In 2018, Orasan attended the 'Academic Research and Innovation Leaders' programme, created to enhance participants' abilities to maximise EU opportunities pre-Brexit and develop strategic thinking for EU/R&I funding post-Brexit.

RGCL staff also benefit from the University's staff development programmes for PhD supervisors and ECRs, research mentoring schemes, and external funding workshops; staff present their work to colleagues at cross-faculty annual research conferences. Research output and development is a key factor in the appraisal of RGCL staff. Standards of research quality and integrity are maintained through training (Section 1.6), and RGCL researchers are offered mentoring to ensure that they receive all support necessary for carrying out their research. The University also provides funding for new academic staff, to enrol on the Post-Graduate Certificate in Higher Education and Professional Practice (PGCert) course.

The ECR development programme provides 30 workshops annually and drop-in sessions to support ECRs in the preparation, delivery and dissemination of research. RGCL also provides ECRs with opportunities to share their expertise through seminars, teaching, and PhD supervision. The ongoing policy of offering mentorship and encouragement to publish has developed successful young researchers capable of pursuing senior research roles. The Doctoral College also provides comparable training and support for our PGRs.

RGCL also provides opportunities for collaboration with eminent researchers and involvement in projects for all staff. An ample conference fund enables all staff to attend two or more international events (conferences and summer schools) each per year. In the REF period, GBP45,000 was provided to support attendance at 43 conferences by 12 staff and 41 students.

Annual appraisals by line managers prioritise staff members' career progression, research development, publications, ongoing funding bids and any potentially beneficial training. These are reviewed regularly throughout the year. RGCL holds workshops in which staff and students can present current research (including research papers and work-in-progress project proposals) to obtain feedback from the group. These workshops are open to colleagues, research students, and the wider community, and are chaired by senior RGCL members. RGCL also collaborates with the University's Project Support Office (PSO) and has two dedicated in-house Project Officers, enabling greater support for academic staff when developing bids and implementing research projects.

Impact is a core strand in the development of both academic and administrative staff within RGCL. The integration of impact within research and teaching is a key focus of several training sessions that are regularly attended by academic staff and RGCL's project officers.

RGCL staff are encouraged to help shape University policies through active participation in the University's ethics and researcher development sub-committees (Evans and Orasan, respectively). Evans helped develop the University's guidance 'Toward Ethical Research Using Datasets of Illicit Origin'.

RGCL also supports requests for flexible working. Since 2018, several staff have requested reductions in, and changes to, their working hours (e.g. Harper 0.8 FTE, Wilson 0.6 FTE), which have been granted. This inclusive approach has enabled RGCL to retain experienced administrative staff and maintain continuity. The current team has served since 2017. The RGCL administrators also have access to, and engage with, university-wide staff training provided by Organisational Development and specialised training. They attend continuous personal and professional training and development, including training in project management, positive psychology, core strengths, and supporting research students.

2.2 Research Students

During this REF cycle, 12 PhD students were enrolled at RGCL. In the previous cycle, eight doctoral degrees were awarded by the unit, and in the current cycle, 10. This indicates RGCL's success in fulfilling its strategic Aim 6 and achieving increased sustainability and development, thus further enhancing RGCL's vitality. PhD students are funded internally (by the University) and externally (funded projects, non-UK scholarship schemes, or self-funding).

RGCL's PGR strategy provides funding for PhD studentships (both fees-only and full) through links to externally-funded projects and internal funding, enabling the unit to recruit the most promising candidates. All students receive further support from the unit and RGCL's research networks, which extend beyond academia.

Staff with expertise in relevant sub-disciplines of modern languages, linguistics, NLP, and translation technologies supervise RGCL's PhD students. The progress and development of each PhD student are nurtured by their Director of Studies and one/two additional supervisors, who offer related but complementary expertise. Students meet supervisors at least fortnightly to discuss their research. The University's degree subcommittee (RIILP's Research Student Board) conducts monthly and annual monitoring of the students' progress. This monitoring requires students to present recent research findings publicly (seminars, poster exhibitions) and to indicate progress toward thesis completion. PGRs are also invited to attend the unit's various research seminar series.

RGCL's success in securing external funding in the previous REF cycle helped to increase the Group's research capacity in two strategic research areas: *Language Processing for Assistive Technologies* and *Translation Technology*. An important part of this growth was in RGCL's recruitment of seven, and the successful completion of four, PhD students in these areas.

RGCL provides desks and computers for all postgraduate research students and for Master's students engaged in collaborative research with other members of the Group. Research students also have access to the facilities listed in Section 3.2. RGCL members are proactive in engaging students in research and collaborating with them on joint papers. Where possible, research students are involved in relevant research projects, further enhancing their transferrable skills.

All research students have access to additional resources, including collaborative national and international research networks, research partnerships, and NLP tools and resources developed in the RGCL. Alongside their postgraduate studies, most PhD students attend lectures from RGCL's three master's programmes and the additional workshops on programming, deep learning, statistics and LaTeX. They are also encouraged to contribute to seminar teaching.

RGCL's PhD students benefit from interactions with leading visiting researchers, several of whom have contributed to PhD supervision. Its rich research environment means that most PhD students publish before submitting their theses, a positive indicator of future achievements. For example,

21 papers by Ranasinghe (PGR), eight papers by Rohanian, and 10 papers by Taslimipoor (both now ECRs) were published after peer-review before the end of their third years during this REF cycle. These were disseminated at international conferences, with attendance supported by RGCL's conference fund.

RGCL's PhD students benefit from opportunities and services provided by the University's Doctoral College (described in the Institutional Environment Statement) and discipline-specific support from RGCL and the wider research institute, RIILP.

Between 2012 and 2016, RGCL coordinated *EXPERT*, an EC-funded International Training Network which trained future world experts in translation technology. This project funded two staff and two PhD positions at RGCL, and provided PGR students with access to training courses that are usually not available. Internal *RIF1* and *RIF3* funding also provided three PhD Studentships in Computational Linguistics and Translation Technology.

Responses to the Postgraduate Research Experience Survey 2019 indicated that RGCL students were positive about most aspects of their research experience and were unanimously positive about their supervision, the responsibilities of researchers, and their progress and assessment. RGCL achieved 89% overall satisfaction. RGCL students were also positive about the numerous research seminar series at RGCL, offered online in 2020. In the next REF period, RGCL will continue to invest in resources for research and teaching and will maintain its focus on the wellbeing and academic progression of its students.

One indication of the vitality of RGCL's environment for research students is the fact that its postgraduate students have gone on to pursue successful careers in academia and applied areas. Examples include: Aziz (Head of the Probabilistic Language Learning Group, University of Amsterdam); Gupta (Machine Learning Researcher at Apple); Stajner (Chief Research Scientist at ReadableAI; Senior Research Scientist at Symanto); and Taslimipoor (Postdoctoral RA, University of Cambridge).

2.3 Equality and Diversity

RGCL shares the University of Wolverhampton's vision of equality and diversity and seeks to have this represented within its population of staff and students. It commits to ensuring that recruitment and promotion in the workplace are fair, transparent, and mindful of systemic and unconscious biases, aiming to have a balanced representation of society across all levels of the hierarchy. As members of the University, all RGCL staff have access to a range of inclusivity services and networks. RGCL staff take University-accredited equality and diversity training and maintain this with regular refresher courses.

Diversity in the Workplace

RGCL is culturally and ethnically diverse, with members (staff and PGRs) from Brazil, Bulgaria, Egypt, France, Germany, India, Iran, Ireland, Lebanon, Mauritius, Mexico, Pakistan, Poland, Romania, Russia, Saudi Arabia, Serbia, Slovenia, Spain, Sri Lanka, Sudan, Turkey, the UK, and Vietnam. The unit provides accommodations for members with specific learning disabilities, chronic diseases, and cultural sensitivities. This includes considered desk space allocations, flexible working hours, and ergonomic workspaces.

RGCL recognises the importance of gender balance in HE and has developed a plan to promote gender equity and diversity. The Gender and Equality Action Plan outlines goals with respect to protected groups and details the measures and actions pertaining to leadership, recruitment, retention, progression, and the working environment to be implemented to achieve these goals. In this REF submission, four of RGCL's returned staff are women, including one Professor and one Reader (2.4 FTE in total). RGCL endeavours to create a fair, balanced and representative professional environment for all groups.

Diversity in Leadership

RGCL practices Equal Opportunities and assesses prospective staff based on their professional potential. RGCL also encourages early- and mid-career staff members to participate in leadership development programmes. In 2017, Moze was awarded a place on the Aurora Leadership Development programme and support was provided in terms of both time and funding.

Staff at RGCL receive training on the promotion of women in science. Corpas Pastor received the Spanish 'Farola' award which recognises women who are leaders in the field of technology. In its commitment to improve equality and diversity in leadership, RGCL is developing formal plans to improve the gender balance of senior academics within RGCL, monitored through the Workforce Planning Process. Information for equality and diversity monitoring will be recorded in all internal research reports and budget monitoring in the next REF cycle. This will be reinforced by regular collection of qualitative and quantitative data to monitor RGCL's progress.

3. Income, infrastructure and facilities

3.1 Income

RGCL has successfully generated external funding, having secured grants from UK research councils, EC schemes, and US organisations and companies. In the previous REF cycle (2014), the UoA that included RGCL generated external research funding equivalent to GBP333,328 per FTE (GBP3.6 million total). In the current REF cycle, RGCL alone has so far generated almost GBP4.3m of external research funding (GBP441,305 per FTE). In addition to the above amount, more than GBP845,000 of commercial funding was secured for projects addressing the needs of for-profit and non-profit organisations, paving the way for future impact and contributing to its vitality and sustainability.

Thanks to its emphasis on collaboration, RGCL's diverse funding sources during the REF cycle included the AHRC, EC funding programmes, the US National Board of Medical Examiners, University of Vigo, Qatar Research Grants Foundation and the Austrian Academy of Sciences. The Group also provided NLP consultancy services and developed bespoke NLP solutions in several projects funded by commercial partners.

The funding obtained was used to support the unit's excellent research. In the REF period, AHRC-funding, EC funding for Marie Curie Innovative Training Network Actions (*EXPERT*) and for Small or Medium-Scale Focused Research Projects (*FIRST*), and commercial funding from the NBME led to the growth of the unit in terms of staff numbers and the appointment of experts in RGCL's key research areas. The EC and NBME funding supported the extension of the unit's PhD programme through collaborative links and bursaries. The research undertaken in the *FIRST* and NBME-funded projects underpinned RGCL's impact case studies. Funding of the *EXPERT, FIRST*, and NBME projects contributed to the salaries of 12 staff members, including appointments of five new research staff and four new administrative staff.

In the next REF cycle, RGCL aims to further diversify its sources of income, facilitated by support from the University's PSO and access to information from services such as *GRANTfinder*, *Research Professional*, and *Funding Insight*. RGCL's research staff, whose duties include preparation of research funding bids, and dedicated Project Officers also attend local events hosting funding organisations such as the EC (Horizon 2020) and UK research councils. RGCL members regularly attend workshops about new funding opportunities organised by the PSO.

Research funding raised in the current REF cycle has been instrumental in achieving high-quality research outputs and impact. RGCL's two impact case studies are based on externally-funded research projects:

- *Language Technology to Improve Text Accessibility for People with ASD* brought impact to organisations providing services for people with ASD. The funding enabled RGCL to evaluate the software it developed, both intrinsically and from the user's perspective, and to investigate the short-term impact of RGCL's research over the final six months of the project. The project led to open-access research and multilingual datasets.

- *Improving High-Stakes Medical Examinations through Natural Language Processing* developed new tools for computer-assisted authoring of assessments of health professionals' educational attainment. This contributed to improved assessment of the educational attainment of US health professionals.

In addition to those arising from these two projects, many of RGCL's research outputs are derived from externally-funded projects, including *DVC* and *EXPERT* (see Sections 1.2.1 and 2.2).

During the REF period, RGCL also benefited from GBP943,167 internal Research Investment Funding for Capacity Building in Computational Linguistics (*RIF1*) and Translation Technology (*RIF3*, 2017-2020). This was used to expand the unit, with *RIF1* and *RIF3* together funding one Reader, one Senior Lecturer and two Research Fellows and providing three PhD studentships.

3.2 Infrastructure Supporting Research and Impact

RGCL maintains a rich scholarly infrastructure supporting research. It occupies 284 square metres of office space, which includes a departmental library providing access to 470 books and journals on computational linguistics, corpus linguistics, phraseology, translation technology, and lexicography. RGCL also provides its members with advanced computational resources, including server machines with approximately 25 TeraFLOPS of GPU and 500 GigaFLOPS of CPU computing performance.

3.2.1 Organisational Infrastructure

The University supports RGCL in a variety of ways, acknowledging its successful track record. RGCL has its own dedicated Research Support Team, which comprises Research Administrators and Project Officers. The Research Administrators support research students while the Project Officers provide pre- and post-award support including dissemination of new funding calls, bid-writing support, and preparation of budgets. The current team has provided high-quality administrative support to RGCL members for the past four years, building a close working relationship with staff and students and progressively gaining subject knowledge. RGCL also promotes peer-review within the Group, encouraging academics to work collaboratively on a wide range of project proposals and funding bids. The Research Administrators help organise events such as the research seminar series, ensuring a collaborative and integrated approach to research.

RGCL benefits from research support services provided by the University's Research Policy Unit and Doctoral College. The support provided by these departments is coordinated via the University's central Research Hub. The Hub manages the University's JeS Research Organisation account through which research grant applications are submitted and coordinates the ResearchFish system which reports outcomes and outputs to RCUK.

Research links with industry are encouraged and developed with the help of a dedicated impact officer assigned to the unit, the University's Business Solutions Centre, and its Science and Arts parks.

4. Collaboration and contribution to the research base, economy and society

RGCL's activities during the REF period involved extensive collaboration and brought a range of contributions.

4.1 Research Collaborations

Our successful collaborations are underpinned by the freedom research staff enjoy in pursuing personal research goals in line with their interests and by the allocation of resources to five different activities (Sections 4.1.1-5). In this indicative overview of arrangements and support provided for the unit's research collaborations, square brackets are used to indicate [collaborators]

and braces are used to indicate {users/beneficiaries}. Some organisations play {[both roles]}. For-profit organisations are underlined while non-profit organisations are *underlined and italicised*. Educational organisations are unmarked.

4.1.1 Research Visits

Hosted by the unit and potential collaborators, research visits have proven useful in fostering more significant collaborative research. Examples include but are not limited to the following strategic topics:

a) *NLP Methods and Tasks*: Prominent researchers (e.g Gelbukh, Ureña) have undertaken sabbaticals at RGCL, establishing and reinforcing collaborative research.
b) *Corpus-Based Studies*: Ongoing collaborations were strengthened through the unit's provision of research visits for staff and students from the [Catholic University of Leuven], [University of Pavia], [Pontifical Catholic University of Valparaíso], and the [University of Naples "L'Orientale"], collaborating on lexicography and phraseology. In total, RGCL collaborated with more than 12 organisations in the REF period on this topic.
c) *Language Processing for Assistive Technologies*: Responding to the University's aim of enhancing interdisciplinary research, RGCL initiated collaboration with numerous experts working on autism (Dr Jordanova, {[*Central and North West London NHS Foundation Trust, CNWL*]}; Prof Selda Ozdemir, [Hacettepe University]) through a series of research visits. RGCL funded additional networking in Europe to build a larger consortium and develop a successful bid for EC funding. In this project (*FIRST*), RGCL collaborated with two educational organisations ([University of Alicante]; [University of Jaen]), three non-profit organisations ({[*Deletrea*]}, Spain; {[*Parallel World*]}, Bulgaria; {[*Autism Europe*]}, Belgium), one health service provider ({[*Central and North West London NHS Foundation Trust*]}), and two companies ([iWeb technologies], UK; [Kodar, OOD], Bulgaria) to develop assistive NLP tools to convert texts into a more accessible form for people with ASD.

4.1.2 Joint Supervision of Research Students

RGCL provides resources to support the mobility of research staff within the unit and from other institutions to jointly supervise research students, creating new opportunities for collaboration and helping identify new potential end users. One example as part of its research in *Corpus-Based Studies* is the unit's joint supervision of a PhD student with {[*Samsung Toronto AI*]}, which strengthened its network of collaborators in the research field of phraseology.

4.1.3 Organisation of Conferences and Evaluation Campaigns

RGCL's organisation of research events has expanded its networks in several strategic areas and paved the way for subsequent collaborations:

a) *NLP Methods and Tasks*: Mitkov serves as Programme Chair at the biennial RANLP conferences. In this role, he established collaborations with researchers at [Simon Fraser University] on the topic of sentiment analysis.
b) *NLP for Digital Humanities*: RGCL supported collaborative research and the establishment of an evaluation campaign (VarDial) focusing on the development of methods to automatically identify dialects and similar languages.
c) *Translation Technologies*: The unit's provision of resources to support Mitkov as Chair of the Translation and the Computer conferences helped to establish and foster collaborations between RGCL and several educational and commercial organisations. This led to RGCL's coordination of the *EXPERT* training network. The network included the {[University of Sheffield]}, {[Dublin City University]}, {[University of Malaga]}, {[Saarland University]}, {[University of Amsterdam]}, {[Translated]} (Italy), {[Pangeanic]} (Spain), {[Hermes]} (Spain), and four associated partners.

4.1.4 Demonstrations of NLP Tools

In the preparatory stage of its research on *NLP for Technology-Enhanced Learning*, RGCL invested in computer resources which were used to provide demonstrations of software for potential users to experiment with. This helped to strengthen the relationship established through networking at conferences to bring about the collaboration presented in one of RGCL's impact case studies. This collaboration with the {[National Board of Medical Examiners]} (USA) led to the development of automatic methods to improve the assessment of educational attainment for medical licensure, to score patient notes, to predict test-item characteristics, and to generate multiple-choice question (MCQ) distractors. RGCL also collaborated with {[Televic Education]} (Belgium) on research in MCQ test generation.

4.1.5 Dissemination via Keynote Speeches and Invited Talks

Dissemination of research findings by senior RGCL staff helped foster productive collaboration on several strategic topics. The unit's research on *NLP Methods and Tasks* involved collaborations on various topics, established as a result of keynote speeches and invited talks delivered by Mitkov in his role as Programme Chair for NLP conferences. In *Corpus-Based Studies*, RGCL's collaborations were established and fostered as a result of keynote speeches by Corpas Pastor, Hanks, Mitkov, and Oakes. RGCL was able to successfully develop the new research area of *NLP for Digital Humanities* by building a wide collaborative network comprising one commercial and 11 educational organisations. RGCL funded Mitkov's trip to deliver a talk at the {[Austrian Academy of Sciences]} in the preparatory stages of the collaborative APIS project, which developed and applied NLP tools to assist prosopographical research. Mohamed was invited to collaborate with the Doha Institute of Graduate Studies due to his original work in cultural analytics. Collaborators and beneficiaries of the unit's research in this area include the {[Jožef Stefan Institute]} (Slovenia), the {[University of Zagreb]} (Croatia), the {[University of Helsinki]} (Finland), {[Harvard Medical School]} (USA), and the {[Rakuten Institute of Technology]} (Singapore). Several collaborations on the topic of *NLP for Social Media* arose through networking by senior academic staff as a result of delivering invited talks. These include collaborations with educational organisations on topics such as hate speech detection ([Harvard Medical School]) and negation and speculation detection in social media ([Simon Fraser University, Canada]). RGCL's collaborations in the area of *Translation Technology* were established through Mitkov's role as chair for Translation Technology conferences and Corpas Pastor's invited talks on translation and interpreting technology.

In this REF cycle, RGCL's collaborative research contributed to its extensive research output: 216 peer-reviewed academic papers (68%) involved external collaborations. The unit addressed 106 research topics in more than 50 collaborations with individuals and organisations in academia and beyond. Collaborators and users/beneficiaries included educational, non-profit, and for-profit organisations. The high volume of research collaborations undertaken is evidence of the unit's successful strategy.

4.2 Engagement with Key Research Users and Beneficiaries

RGCL collaborates with partners from academia, government organisations, NGOs, SMEs and large commercial companies. It has impacted all these organisations through knowledge transfer and consultancies. RGCL engaged with the key research users and beneficiaries specified in Section 4.1.

RGCL's research on *NLP for Technology-Enhanced Learning* was exploited by NBME, a non-profit organisation which develops and administers the US Medical Licensure Examination. They used RGCL research to improve the quality and efficiency of the services they provide (see: Impact Case Study, *Improving High-Stakes Medical Examinations through Natural Language Processing*).

As part of the *FIRST* project, RGCL engaged with NHS Foundation Trusts, non-profit associations, and software development companies. These users directly applied the unit's research in Language Processing for Assistive Technologies to improve the accessibility of services for people

with learning disabilities and ASD, and neurotypical users (see RGCL's impact case study on *Language Technology to Improve Text Accessibility for People with ASD*).

End users who have engaged directly with RGCL in *Translation Technology* include commercial translation firms such as Hermes, Wordfast and Pangeanic. These users were engaged by RGCL as part of the EC-funded *EXPERT* project, in which NLP technology was developed to improve Translation Memory programs for translators. This involvement of professional translators was so successful that RGCL is now planning similar engagement with the growing number of companies (currently 25) acting as associated partners in the ongoing EMTTI project, several of whom have already hosted RGCL students on placements.

Other end users include the Doha Institute for Graduate Studies on Arabic Cultural Analytics and Qatar Museums. Both beneficiaries have made use of, and given useful feedback on, tools created by RGCL (e.g. a web framework for cultural analytics that enables Humanities scholars to use NLP tools without programming).

4.3 Wider Contributions to the Economy and Society

Wider societal contributions were brought about by RGCL's research on *NLP for Technology-Enhanced Learning* in the project "Towards European Language Learning for Medical Professionals" (*TELL-ME*). In *TELL-ME*, RGCL collaborated with Saarland University and Mannheim University Hospital (Germany), the University of Malaga and Hospital Pascual (Spain), and the NHS West Midlands Workforce Deanery (UK) to develop innovative NLP-based solutions to support the cross-border mobility of healthcare professionals. This included the provision of self-study materials via an innovative e-learning platform for professionals to learn vocational language (English, German, and Spanish) to improve communication with patients. Thus, the project addressed an important challenge in the provision of health services in Europe. RGCL is currently collaborating with the University of Malaga on a project using Neural Machine Translation Technology to facilitate communication of English and Arabic speaking patients in Spanish hospital triage scenarios.

In collaboration with Hacettepe University, RGCL developed deep learning methods to analyse descriptions of Android applications and identify their requested permissions, with the aim of blocking potentially harmful or unnecessary permission requests. As a result, this research has the potential to make a significant impact on cyber security in the Android market.

4.4 Engagement with Diverse Communities and Publics

One of RGCL's main research areas has been language processing for assistive technologies, particularly the development of tools to improve text accessibility for people with ASD. This involved collaborations with several charities and an NHS Foundation Trust to perform interviews, and run surveys and reading comprehension experiments to better understand the specific difficulties faced by people with ASD and their carers. RGCL presented the research underpinning this assistive language technology to both end users and the wider public at a launch event at Stevenson House in London.

Yaneva worked part-time with the charity Autism West Midlands, enabling RGCL to expand its research network and raise awareness of its research activities among the local autistic community (including carers and other intermediaries). This also facilitated the recruitment of participants in eye-tracking experiments to learn more about the reading behaviour of people with ASD.

4.5 Contribution to Sustainability of the Discipline

RGCL's main contribution to the sustainability of NLP is its emphasis on applied research and user-focused evaluation of NLP solutions. This includes the use of innovative evaluation methods and extrinsic evaluation of translation technologies and assistive NLP applications. The Group has

conducted seminal work on anaphora resolution, automatic generation and evaluation of multiple-choice tests, new-generation translation memory systems, methods for quality estimation in machine translation, and computational phraseology. RGCL academics developed a range of linguistically-annotated corpora and datasets which serve as gold standards for evaluation of NLP systems for human and machine translation, identification and translation of multiword expressions, dialect identification, offensive language detection, and human reading behaviour. RGCL's development of PDEV is a major contribution to computational lexicography.

4.5.1 Support for and Exemplars of Interdisciplinary Research

As noted in Section 1.4, Computational Linguistics research is inherently interdisciplinary. RGCL's applied research in language processing for assistive technologies is underpinned by psycholinguistic experiments. To support its interdisciplinary research in this area, RGCL acquired eye-tracking equipment and funded attendance at psychology and web accessibility conferences.

Staff have also adopted state-of-the-art neural methods from AI and applied them to new areas such as machine translation quality estimation, partial syntactic parsing, and shallow syntactic analysis.

RGCL regularly invites scholars and translation/interpreting practitioners to undertake collaborative visits.

4.5.2 Response to National and International Priorities and Initiatives

RGCL led the *EXPERT* Innovative Training Network in response to the EU priority to equip researchers with the necessary skills for successful public- and private-sector careers. It provided research and training in translation technologies.

RGCL's *FIRST* project responded to the EU priority for European organisations to cooperate in the development of advanced ICT-enabled solutions "for the empowerment of people with disabilities who are at risk of social exclusion" as a result of "low literacy" resulting from cognitive and mental impairments.

4.6 Indicators of Wider Influence, Contributions to and Recognition by the Research Base

RGCL's contribution to the research base during the census period includes its chairing and/or organisation of the biennial International Conferences 'Recent Advances in Natural Language Processing', including the RANLP Summer School on Deep Learning in NLP (Mitkov); the annual Translation and the Computer conference in London (Mitkov); the SemEval-2019 shared task on Identifying and Categorising Offensive Language in Social Media (Zampieri); and 21 other conferences. Members of the unit have also served in a variety of editorial roles: Executive Editor of the Journal of Natural Language Engineering, Cambridge University Press (Mitkov); Editor-in-Chief of the Natural Language Processing book series from John Benjamins (Mitkov); Editor of the forthcoming second edition of the Oxford Handbook of Computational Linguistics, Oxford University Press (Mitkov); Associate Editor of the journal ACM Transactions on Asian and Low-Resource Language Information Processing (Can), 32 other editorships; and 190 editorial board memberships. Several members served as peer reviewers for the Journals of Computer Speech and Language (Mohamed); Natural Language Engineering (Can, Evans); International Journal of Lexicography (Moze); International Journal of Data Mining Science (Can) and were involved in 56 other cases of journal-paper reviewing; and 187 cases of reviewing for research conferences. RGCL staff served as external examiners of 28 PhD students.

Its open-access resources (Section 1.5) represent another part of RGCL's contribution to the research base. Through its Master's and PhD programmes, RGCL has helped to develop young researchers in relevant sub-disciplines of modern languages and linguistics.

Recognition of RGCL's standing is demonstrated by the large number of keynote speeches and invited talks delivered by its members during the census period. Prominent examples are Mitkov (61 in total), Hanks (30) and Corpas Pastor (21). In recognition of his outstanding contributions, Mitkov was awarded the title of Doctor Honoris Causa at Veliko Tarnovo University in October 2014.

During the REF period, RGCL staff have received best-paper awards at numerous conferences (e.g. Can at the 3rd Workshop on Representation Learning for NLP; Yaneva at Web4All 2018, 2019, and 2020, EUROPHRAS 2017, and RANLP 2017) and have developed award-winning systems in competitive shared tasks.

Hanks's corpus pattern analysis (CPA) technique, which underpins his Theory of Norms and Exploitations, has been widely influential, inspiring several lexicographic projects for languages other than English, including Italian, Spanish, Croatian, and Dutch. CPA has also been applied to discourse analysis, subtitling, machine translation evaluation, and language teaching.