IntelliText - Intelligent Tools for Creating and Analysing Electronic Text Corpora for Humanities Research

Research Context

Much humanities research relies on or would benefit from analysis of electronic corpora – representative collections of texts (such as books, newspaper articles, technical manuals, legal documents in computer-readable format), which may also be annotated with linguistic or domain information. The main advantage of using corpora over hand-picked examples is the ability to collect data systematically, to assess the centrality of certain features to the research material, and to establish experimentally potential trends in the data. Modern electronic text corpora often contain over 100 million words, but specialised corpora can be smaller, e.g., only several thousand words). Projects which rely on electronic corpora can be expected to have greater academic and social impact, thanks to increased consistency in data analysis.

However, the major difficulty faced by corpus-based studies in humanities research is that creating and annotating a new corpus and designing an appropriate search engine for textual analysis require complex technical support, e.g., expertise in programming, database management, web development, as well as knowledge of specialised tools for data annotation and analysis. Such a level of technical expertise is often unavailable to smaller humanities projects; but even larger corpus-based projects often miss opportunities for data analysis because of inadequate methodological or technological support for relevant computational aspects. Even when a corpus already exists, the task of building appropriate computational tools for analysing, intelligently searching and visualising the data still remain too challenging for many potential humanities projects.

Humanities researchers' lack of awareness of modern computational techniques for corpus-based studies can seriously limit the scope and the impact of any planned research projects. Moreover, computer scientists who design corpus-based tools frequently do not understand the specific needs of humanities research; their tools are often difficult to adapt to a specific project, or lack an intuitive interface and documentation. As a result, important potential synergies for the research of both parties have been neglected, in particular the applicability of several non-trivial computational techniques to preparing and analysing corpus data, with the power to reveal new dependencies and patterns in the material, and thus yield a much greater impact.

IntelliText aims to create software to allow humanities researchers with no specialised background in computer science or corpus linguistics to take advantage of advanced methods of text collection and analysis. It will enable them to collect new project corpora from the web, have them enriched automatically with linguistic and other annotations, and then easily uncover interesting patterns of usage, starting either from their own intuitions and hypotheses, or from expressions and patterns identified as potentially noteworthy by the system.

The software will be designed and tested in conjunction with target applications in three areas: translation studies, language teaching and monitoring opinion and sentiment. These will demonstrate its generalisability for addressing the needs of a wide spectrum of humanities researchers, including historians and literary specialists. IntelliText will offer an intuitive and well-documented interface and be made freely available for research purposes.

Added Value

IntelliText will not address issues related to the content of existing electronic corpora, digitising paper-based documents, or building particular general-purpose corpora. Rather we will concentrate on enabling **intelligent access to researcher-created corpora**. Thus, our work focuses on the first aspect of the AHRC DEDEFI call – Enhanced access to Digital Technologies for innovative research in the arts and humanities.

IntelliText's main advantage to humanities researchers across the piece will be the ability to collect and analyse textual data in ways that are currently beyond their reach. Researchers in contemporary history, politics and media studies, for example, will be able to maintain and analyse

a *monitor corpus* – a collection of electronic texts enlarged on regular basis, e.g., a collection of newspaper articles or blogs published daily on selected websites together with reader commentaries. These can be automatically enriched by explicit annotation of affective expressions and also people, organisations and geographical locations. Such a resource enables projects aiming to monitor trends in public opinion and perceptions, or analyse shifts in political rhetoric and language, allowing researchers to build frequency time-lines of specific words and expressions to substantiate their conclusions.

As a further example, for researchers in translation studies IntelliText will automate: the downloading of large collections of translated texts and the originals from the web in the languages, subject domains and genres their projects require; the annotation of each word with part-of-speech and base form; and the alignment of the texts sentence-by-sentence and (where possible) word-by-word, so that the alignment becomes an additional feature for identifying interesting usage cross-lingually. In particular, this resource will be useful for research in *translational stylistics*, which attempts to determine the characteristic features of translated texts as such, focusing on, for example, sentence-initial elements or cohesive devices like conjunctions.

Researchers in any humanities area will be able to work with the whole range of electronic genres beyond literary texts, such as analytical or informative newspaper articles, technical manuals, sports reports, reviews, and blogs.

IntelliText's novel contribution will be to tune advanced tools and methods from computer science to the needs of humanities researchers, integrating them into a single software application with a simple interface and good documentation. The majority of these tools are already available as separate systems, e.g., for automatic part-of-speech annotation, annotation of proper names, alignment of translated texts, term extraction, identification of synonyms and related terms. By introducing these tools and methods into fresh research areas, we anticipate that synergies with specific corpora and research goals will lead to further extensions of the software. IntelliText will put into the public domain and distributed as open-source software to academic and industrial users, who will be free to extend it under the same conditions for the benefit of the research community beyond the initial funding of the project.

To summarise the impact of IntelliText, it can strengthen the theoretical foundations of many humanities disciplines by enabling a much larger community of researchers than hitherto to make testable predictions about their subject, and then to verify (or falsify) them by reference to solid corpus evidence uncovered by advanced and automated analytical techniques.

Quality, Innovation and Track Record

Our main aim is to create a software system to support corpus-based research across the humanities and to evaluate its usefulness within three sample research applications. The system will address the methodological needs of researchers by automating the creation of their own project-specific electronic text corpora, and their subsequent annotation, analysis and visualisation. IntelliText will exploit potential synergies between existing computational methods and tools for corpus processing and the content of the sample projects. The project will deliver:

- an open-source software package which integrates existing methods and free tools for the automatic collection, annotation and analysis of electronic text corpora (as specified in the *Technical Annex*);
- comprehensive documentation on downloading and using the software for each task of the corpus-based research;
- case studies using the software within three application domains, described in *Potential Usage*.

The Centre has an internationally outstanding track record of creating collections of texts from the Internet in a variety of languages (Sharoff 2006), classifying texts by their domain and genre (Sharoff 2007; Kurella et al. 2008a), estimating their difficulty (Sharoff 2008). From early 2010 the Centre will consolidate its standing with leading European research and user groups in developing corpus software within two FP7 projects on machine translation and an EU Lifelong Learning

project on extracting language-learning resources from multilingual corpora. IntelliText will adapt and extend tools previously created by us in the context of the Ars Rococo project (Ciobanu et al. 2006), funded by The British Academy, and ASSIST, funded by EPSRC (EP/C005902).

In addition to working with colleagues from complementary academic disciplines, we have developed a track record of collaborating with colleagues from industry to reach a wider community of beneficiaries. In a three-year project with the Translation Automation Users Society, the Centre is adapting the ASSIST software to provide open access to a one-billion word repository of technical translations in 84 language pairs, while Google is funding research on the automatic classification of web genres. The ProLingua project delivers tailor-made brand tone of voice audits for clients from the financial services sector, based on corpora of around 500k words for each brand, including pages from corporate websites and printed documents. Thus we can distinguish qualities of a brand, such as "reliable" or "dynamic", on the basis of linguistic evidence, identify inconsistency in a given set of documents, or investigate patterns of variation in communications for different audiences. The corpus-based approach places our analysis on a much sounder empirical footing than considering smaller numbers of hand-picked examples. We can distinguish those features which occur with a significant frequency and those which are perhaps anomalous. Other analytical approaches risk exaggerating the importance of the latter. In sum, our approach allows us confidently to tell brand owners things about their communications of which they were themselves unaware and to suggest specific areas for improvement.

Potential Usage of the Software

IntelliText will be developed and evaluated in concert with other researchers engaged on projects in three application areas, and who use text corpora as their main source of data. We will elicit their needs and their feedback on the specifications of the system, and involve them in formal evaluations of both functionality and usability.

1 Translation studies

Ongoing projects on evaluation in translation and the contrastive analysis of semantic prosody are led by Munday (2009; forthcoming a; forthcoming b). The first project is looking at linguistic markers of translator stance (evaluative epithets, modals, connotative lexis, etc.) and relates to corpus-based work on evaluation in English media discourse (Bednarek 2006[1]). The project currently focuses on the qualitative, manual analysis of translated texts (literary and political) from French and Spanish. We will work with Munday to ensure IntelliText facilitates a more systematic analysis of such lexical items and their various translations in large corpora, thus considerably expanding the scope of his work. We will collect and annotate a representative corpus, and display aligned search results at the level – text, sentence, phrase – that best assists the analyst.

Work in contrastive semantic prosody traditionally relies on large representative linguistic corpora (like BNC or the corpus of the Royal Spanish Academy), but genre-specific data would be more informative. IntelliText will be tuned to collect monolingual and parallel corpora specifically for those genres which are under-represented in such general collections. We will also adapt advanced techniques for automatically identifying semantically-related words within bilingual corpora (Sharoff et al. 2009) to explore this novel research question. .

IntelliText has the capacity to further stimulate translation theory by supporting future projects in, for example, translational stylistics, translation of cultural elements, proper names or metaphors, analysis of difficult-to-translate expressions (cf. Corness 2009; Butler 2008). In particular, it will open up many PhD projects currently impractical for want of corpus resources.

2 Corpora for language teaching

The software will be tested within a Computer-Assisted Language Learning project being conducted at the Centre for Translation Studies, Leeds. (Kurella et al., 2008). It aims to create collections of recent, topic-related texts for language learners. The project will use all major functions of IntelliText and can therefore be used for its calibration and testing. In particular, we will

[1] See Technical Annex.

create a monitor corpus of newspaper articles of different types in several languages, annotate them for various features and use this annotation to select texts suitable for classroom use that exhibit particular grammatical structures or words from specified topics, which can be highlighted in the text for presentation to students.

Aligned parallel texts collected and annotated by the system will be used for creating translation exercises. Teachers will be able to hide or selectively display published translations of selected texts, collect students' translations and automate a partial comparison of student and professional versions to prompt further discussion and analysis.

In general, IntelliText will facilitate discovery learning of contrastive features between languages, where particular linguistic constructions (e.g., passive sentences) can be quickly selected from corpora, and compared with similar types of constructions in the other language – in terms of frequency, context of usage and distribution.

3 Opinion and sentiment monitoring

In media research the corpus processing environment can be used to maintain monitor corpora of the most recent news or analytical articles in order to investigate shifts of opinions over time, signalled by the use of key words and expressions similar in principle to those targeted in the problem domains of evaluative language in translation and tone of voice in customer relations. Since history took a cultural and linguistic turn over 20 years ago, many of the uses suggested for scholars of languages, linguistics and media studies would also be of interest to historians. For example, print media historians will be interested in the uses suggested for contemporary media analysts. The system would be equally useful for historical analysis of past periods to show key cultural shifts. Literature researchers may enjoy a functionality that can identify similar linguistic contexts in texts in order to systematically search for reminiscences and influences between different authors, and even literary epochs.

In this application domain we will re-use techniques developed by Prolingua. For example, adjectives often perform an evaluative function in language and so frequency lists for adjectives and adjective-noun collocations can provide a means of identifying the stance of the speaker, as can an analysis of the contexts of verbs indicating mental or verbal processes.

Sustainability and Reusability

The system will be hosted on the Leeds University server system and will be maintained after the end of the project by the Information System Service of the University. The software package will be uploaded to a community software website (*Google Code* or *SourceForge*), and in this way it will be maintained by the research community beyond the lifetime of the project: contributors to our open-source project from around the world will be able to contribute their time and skills to further improve the software and extend its functionality. This model will allow us to ensure that the system is well-maintained and continues to evolve with input from multiple users within well-maintained and free community software infrastructures.

Project Management

IntelliText will have a Project Board, consisting of leading humanities researchers. The Board will meet quarterly to re-assess vision, priorities, progress, and to promote academic and industrial uptake of the project.

The following have agreed to be Board members:

- Prof Tony McEnery, Lancaster (corpus linguistics, computer-assisted language learning)
- Dr Jeremy Munday, Leeds (theoretical translation studies)
- Prof Rob Waller, Reading (language simplification for corporate and government communication)
- Prof Simon Burrows, Leeds (modern European history)

- Prof Clive Upton, Leeds (dialectology)

The project is carried out by the Centre for Translation Studies, School of Modern Languages and Cultures, University of Leeds. The project investigators are Prof Tony Hartley and Dr Serge Sharoff.

The Centre was created in October 2001 as an interdisciplinary research centre with a particular focus on the development and evaluation of technologies designed to support translators, interpreters and subtitlers in their work. In RAE2008 CTS researchers joined with the AI Group in the School of Computing, achieving GPA 3.05 (4* 25, 3* 55, 2* 20). Externally, its academic collaborators include Lancaster, Sheffield, Athens (ILSP), Barcelona, Nantes, Romanian Academy of Sciences, Saarbruecken, Stuttgart, Tokyo and Zagreb. The Centre enjoys the support of the School of Modern Languages and Cultures, the largest and linguistically most diverse of its kind in the UK, with a staff of over 140, over 200 postgraduates, recent SRIF investments of some £945k and a total annual budget in the region of £14 million.

Tony Hartley is Professor of Translation Studies at Leeds. Current work focuses on using corpora in several languages, domains and genres both to extract equivalent terms and expressions, and to develop automated metrics for evaluating the quality of machine translation output, with results published at ACL, COLING, EACL, LREC and MT Summit. He has extensive experience of generating e-learning resources for training translators in the use of corpora and language technologies and disseminating them through the websites of the eCoLoRe, eCoLoTrain and current eCoLoMedia EU-funded projects. He previously worked on and managed a number of UK (EPSRC, DTI) and EU projects on language technologies, including ASSIST, DRAFTER, AGILE, FABULA, EUROMAP and TIC. He has held visiting appointments in Sydney University's Computer Science Department, at the Communications Research Laboratory, Japan, and Tokyo University.

Serge Sharoff is Lecturer in the Centre for Translation Studies. He was the leader in development of the Russian National Corpus (Sharoff 2005), comprising 100 million words covering a wide variety of types of texts in modern Russian. Subsequently he developed a technology for automatic acquisition of representative Internet corpora (Sharoff 2006), resulting in a collection of corpora of 100-200 million words for English, Chinese, French, German, Italian, Japanese, Polish, Portuguese, Russian, and Spanish, as well as tools for their classification (Sharoff 2007).

We will employ two post-doctoral researchers, who will be responsible, respectively, for software integration and for liaison with researchers in the target areas. Their responsibilities are specified in the *Technical Annex*, which also describes the workpackages shown in the Technical Annex

Dissemination

We will host a one-day workshop in Leeds in month 1 of the project with the Project Board and 10 other academic and industrial leaders to brainstorm on requirements across the humanities and establish outreach channels. We will publicise the project on the AHDC website; the software will be made publicly available via Leeds University servers and free community software websites (like Google Code). On these websites potential users will also contribute their ideas, code and documentation starting from the initial stages of the project: planning and defining specifications of the system. Via these websites we will monitor contributions and uptake of the system (e.g., geographical distribution of downloads, projects using it) beyond Leeds.

We will present our results at the following conferences:

- Translating and the Computer conference (London, November 2010)
- Teaching and Language Corpora conference: TaLC 2010 (Europe, July 2010)
- Language Resources and Evaluation Conference (LREC) (Europe, May 2010).

Word count ~3000 words