# IntelliText – Environment for the creation of and intelligent access to electronic text corpora: a novel platform for humanities research
# Impact Plan

The proposed project will address the needs of the corpus-based research community both in academe and in industry, and in the longer perspective its results will be used to improve corporate and government communication.

The IntelliText project relies on established technologies developed in Computational and Corpus linguistics; however these areas will not only contribute their Open Source tools and methods to our project, but will also benefit from it. Research in computational linguistics recently gave rise to *linguistic engineering*, which uses advanced methods in speech and language processing to build commercial computer systems that work with texts or with speech, or which embed this technology into intelligent robotic systems, bringing it within the wider field of Artificial Intelligence.

We expect that the IntelliText project will have a direct industrial impact in the area of linguistic engineering and speech and language technology: the deliverables will be useful for industrial companies which build software system for machine translation, speech recognition, text-to-speech, information extraction, and automatic question answering. Nowadays such technologies rely on large annotated corpora of electronic texts for development and evaluation of the software. However, the quality of such systems is limited by the quality of the corpora used. For such industrial users the IntelliText project will deliver a platform for targeted, fine-grained collection and annotation of corpora For example, the system will integrate state-of-the-art tools that can harvest corpora in a specific subject domain or genre. These corpora can greatly enhance the quality of the linguistic engineering technologies designed for specialised domains, since the data then becomes cleaner and less ambiguous.

In addition, the IntelliText system will contain a coherent set of tools that work together, which will save much time and effort for industrial users by automating the most common workflows and tasks, like collecting, annotating and aligning parallel corpora. This aspect will be especially important for small linguistic engineering companies and start-up projects, who cannot dedicate large resources to building and maintaining their own software infrastructure.

Linguistic analysis of tone of voice for corporate branding and text simplification for corporate and government communication are other areas where the IntelliText system is expected to make an industrial impact. The proposed technologies will improve branded communications in this area by automating key stages for collection, annotation and – most importantly – analysis of the corpus-based linguistic data. The system will enable language analysts to view the data from multiple perspectives and to make more accurate conclusions on connotations and expected effect of the language used in corporate publicity materials.

Similar technologies will be useful for opinion mining and sentiment monitoring, which nowadays have important commercial and social applications. For example, companies which are interested in accurately monitoring public opinion and uptake of their products will be able to monitor large collections of diverse sources such as blogs, reviews and analytical articles, internet discussions in dedicated forums with a view to discovering specific features and functions of their products which need to be improved or added in future releases, if they are to stay competitive. The IntelliText system will enable such automatic mining of opinion via collecting a monitor corpus from specific sources and focussing on specific words and expressions used in this corpus.

The professional translation community will also benefit from using the proposed system. Nowadays the translation industry makes extensive use of computer-assisted technologies, such as translation memories, electronic terminological databases and dictionaries, and post-editing of machine translation output. One common feature of these technologies is corpus-based analysis of how linguistic expressions are used in the source and target languages. Professional translators nowadays use large monolingual and bilingual corpora to analyse this usage, but there is a growing need for specialised corpora (domain-specific or even user-specific), where translators can look for terminology or linguistic patterns.

By using IntelliText translators will be able to collect and query data for their own projects within subject domains and genres of their areas of interest. They will be able to study networks of related terminology built automatically from such corpora, in order to better understand subject domains and the usage of terms within the terminological systems. They will also be able to map these systems across languages using existing parallel corpora and find translation equivalents more efficiently. Finally, they will get support for translation across different underlying domains. For example, if a foreign legal system is different from the UK system, then terms in the two languages will not exactly map to each other; the system will be able to identify maximally similar translation equivalents and highlight the need to provide necessary explanations. Commercial companies that specialise in computer assisted translation will be able to build specific applications for translators which facilitate such exploration of subject domains and terminological systems across languages.

In order to ensure that the project reaches users beyond academia, we will present the system at conferences with industrial participation (Translating and the Computer, Language Resources and Evaluation Conference – LREC2010), and we will also promote the system on the knowledge transfer section of our University of Leeds website.

The Centre for Translation Studies (CTS) at the University of Leeds has established long-standing relationship with several industrial partners in the areas of translation technologies and linguistic engineering. In particular, since 2008 the Centre has been a partner in a three-year knowledge-transfer project with the Translation Automation User Society (TAUS). TAUS (http://www.translationautomation.com/) is a not-for-profit community of users and providers of translation technologies and services, whose members include Adobe, eBay, Intel, McAfee, Microsoft, Oracle and Sun and other international organisations and companies that daily generate large volumes of documentation in many languages. TAUS has created a fellowship in CTS to support research and development of Intelligent Access to Translation Resources: large collections of translated texts (currently over one billion words), aligned sentence-by-sentence for 84 language pairs.

In 2010 CTS starts two EU-funded projects within Framework Programme 7 ICT Call 4 for “Small or medium-scale focused research projects” (STREP) together with several academic and industrial partners from EU countries. The projects (ACCURAT and TTC) have a strong emphasis on the knowledge transfer component and focus on developing new technologies with several industrial partners: Syllabs SARL (France), Tilde SIA (Latvia), Linguatec (Germany), Zemanta (Slovenia).

For our IntelliText project we will approach TAUS and our FP7 industrial partners with a view to using the developed system for streamlining workflow for creating linguistic resources. In this way different functions of the software may be tested for large-scale development projects. This experience will be publicised on the project website, which will attract other commercial companies working in the field.

For the industrial research community we will publicise the system's functionality, integration of different components and availability of documentation, which will result in a considerable reduction of the development and data analysis efforts, and make it possible to:

- download and clean corpora from the web automatically in a targeted way for R&D projects
- annotate the data on multiple layers, and automatically align translated texts
- use an advanced search engine for presentation and analysis of the data for R&D, e.g., to build new linguistic resources for software applications
- extract terminology and build networks of synonyms and related terms
- view data from different perspectives in the user interface.

In our communications with industry we will emphasize the benefits of these features specifically for commercial R&D projects and for corporate and government communication, giving examples of successful projects developed in conjunctions with IntelliText.