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

Background

There are many collections of electronic texts available for humanities researchers. Large repositories of electronic resources "which result from research and teaching in the arts and humanities" are hosted by the Arts and Humanities Data Service (http://www.ahds.ac.uk/), which includes the Oxford Text Archive (http://ota.ahds.ac.uk/) – large collections of electronic texts "for use in Higher Education, in research, teaching and learning".

Our project is intended to enable humanities researchers to create and exploit their own, project-specific corpora and thus deals primarily with automated tools for collection, annotation, analysis and presentation of such electronic corpora to users. These tools usually accompany corpora created for research in linguistics, modern languages, and translation studies. Corpora created for these areas usually contain some elaborate search mechanisms, which work with multiple annotation layers (like part-of-speech, headwords, syntactic structure) and can generate representations for linguistic queries, such as concordance lines in the "Key Word in Context" (KWIC). Interfaces to these corpora also implement certain data analysis methods which can summarise the usage of linguistic expressions, e.g., by creating lists of *collocations* for user queries (words which are more often used together with the given word of phrase).

Large linguistically-oriented corpora available on-line usually give access to static, pre-defined text collections, such as the British National Corpus (BNC) (http://www.natcorp.ox.ac.uk/), Collins Wordbank's *Online* English corpus (http://www.collins.co.uk/Corpus/CorpusSearch.aspx), Bank of English at the University of Birmingham (http://www.english.bham.ac.uk/research/language/corpus.shtml), InterCorp (http://ucnk.ff.cuni.cz/), or our multilingual Leeds Internet corpus (http://corpus.leeds.ac.uk/internet.html) and our interface to the Europarl corpus of EU parliamentary proceedings (http://smlc09.leeds.ac.uk/eudemo/). These resources are extensively used by researchers for analysis and as a source of evidence about the usage and frequency of specific words or multiword expressions. However, these resources are not designed to collect the data for individual user-defined projects and to query these dynamically-created collections. They also have fixed annotation, which for on-line corpora cannot be extended or modified. The way data is presented to the user cannot be modified either. This seriously limits the usefulness of corpus-based resources for many humanities research projects. For example, annotation of paragraph boundaries is not accessible in the corpora mentioned above. However, this information is essential for analysis of textual cohesion: the place of connectives in the beginning or in the middle of a paragraph often signals the type of their discourse function. Also, many corpora do not offer access to the complete texts from which concordance lines are extracted, and this undermines their usefulness for discourse research projects and for many language teaching tasks, where students need to work the text as a whole.

Sketch Engine (http://www.sketchengine.co.uk/), which primarily targets lexicographical research, goes a step further and allows users to collect their own corpus, upload it and access it on-line. However, users have little control over the way their corpora are collected: the system automatically queries Internet engines like Google or Yahoo with user-defined words and word sequences from a list, so users cannot distinguish different text types or genres which get into the collection; they also cannot download corpora from pre-defined websites or RSS feeds. This system offers no support for annotation of user data, nor does it allow any flexibility in the way the data is presented to the end user. Researchers have no way to discover and analyse interesting features in their corpora, like terminology or proper names; and it does not align or represent parallel texts (original texts and their translations into other languages). Finally, Sketch Engine is not free even for academic use, which is another limitation on its usability for research projects.

The Prototype Text Analysis Tool available at TAPoR Portal (http://taporware.mcmaster.ca/) allows users to analyse the data on specified web pages and to upload user files. It presents the results

either as word frequency lists or as concordance lines in KWIC format. However, technically this tool works with individual texts, since it does not recognise text boundaries in uploaded files, and it does not work with multiple webpages from a website. In this respect, the data analysis capabilities of the system are limited. For example, it cannot distinguish words which occur in almost every text in the collection from those words which are only used in a small number of texts, yet this information may be important for distinguishing content words and function words, discovering terminology, identifying shifts in usage and for other related tasks.

There are several corpus processing tools for off-line usage, like MonoConc, ParaConc (http://www.athel.com/), WordSmith (http://www.lexically.net/wordsmith/version5/index.html), which target primarily linguistic research. They query user corpora, but again they are not flexible on how the results are presented to users (they do not display paragraphs or entire texts),;they do not support user annotation beyond part-of-speech tags and headwords; and they do not help with collecting, aligning, annotating corpora or analysing data for interesting properties and features like terminology, multiword expressions, word time-lines, automatically computed synonyms and related terminology, similar or related texts.

On the other hand, there are free tools for linguistic annotation and analysis of electronic texts, but they are often difficult to use, lack intuitive interfaces or clear documentation. As a result, some technical background is necessary even to understand what their potential benefits for humanities projects might be. For example, Corpus Workbench is used to represent large corpora with complex annotation (http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/), while our CSAR package (http://csar.sourceforge.net/) is a web interface to the Corpus Workbench engine. TreeTagger (http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/) automatically creates part-of-speech and headword annotation for several languages; the SemanticVectors package (http://code.google.com/p/semanticvectors/) automatically creates lists of synonyms and related terms; Giza++ (http://code.google.com/p/giza-pp/) and HunAlign (http://mokk.bme.hu/resources/hunalign) align sentences and words in parallel texts; GATE ANNIE (http://gate.ac.uk/ie/annie.html) annotates proper names. All such tools would be very useful for many research projects in humanities, but they their installation and usage is not intuitive and many of them do not store the data in the same format and therefore require specialised conversion programmes to work together. As a result, researchers rarely use such tools for their projects in humanities simply because of prohibitively high requirements of technical competence, the long time needed to fix incompatibilities and the lack of user-friendly documentation.

At present there is no integrated system which could download and automatically annotate new electronic corpora and supply a flexible and user-friendly mechanism for intelligent access, analysis and presentation of the data for a broader range of research projects in the humanities.

Methodology

During the project we will create a system that will automate the technically-intensive tasks,which researchers typically need to perform in the process of building and providing intelligent access to electronic text corpora.

These tasks will include:

- **Downloading corpora from the web automatically**: This will be achievable both in a targeted way (from websites and RSS feeds specified by the user), as well as in unrestricted way (based on queries to internet search engines) We will use our implementation of the Leeds BootCat technology (Sharoff 2006a) and will extend its functionality with new functions for RSS and recursive website downloads. Further extension will support user-defined filters for downloaded electronic texts, which will allow users to collect corpora by certain features, e.g., texts containing words from a specific list or on a particular topic, or containing names of certain people and places. The system will also support the possibility of maintaining a monitor corpus – to trace specific websites or RSS feeds on a regular basis and extend existing corpora by downloading new, recently published texts. This functionality will be important for media studies or for language teaching projects, where using fresh teaching material is important. The system will enable the preservation of time stamps for downloaded texts, which may be useful for building time-lines of linguistic expressions or proper names.

- **Cleaning collected corpora in a flexible user-defined way**: In particular, the system will convert html files into plain text, but if necessary will preserve information about specific features required by users, such as information about chapter titles, section headings, paragraphs, additional text boxes, external links to other texts in the collected corpus. For this task we will integrate and extend our corpus cleaning tools developed during the CleanEval exercise (Baroni et al, 2008).
- **Annotating corpora with part-of-speech tags and headwords (lemmas) for several languages**: This will be enabled for English, French, German, Spanish, Portuguese, Italian, Russian, and Chinese. Annotation of proper names and syntactic structure will be enabled for English. To do this, we will integrate free corpus annotation tools which work with many languages, such as TreeTagger (Schmid 1994) and GATE ANNIE (Cunningham et al. 2002).
- **Automatically aligning parallel texts**: Original texts and their translations into another language will be alignable at the levels of paragraph, sentence and, where possible, individual words. For this task we will integrate free language-independent alignment tools: HunAlign (Varga et al. 2005) and Giza++ (Och and Ney 2003).
- **Customising a fast and flexible search engine for electronic text corpora**: This needs to work with multiple layers of user annotation, e.g., annotation of individual words and of phrases, sentences, paragraphs, complete texts, etc. For this task we will use the free Corpus Workbench system (Christ 1994) that is the core engine behind our Leeds internet corpora.
- **Automatically discovering and annotating terminology and multiword expressions**: For this task we will use techniques and software developed during our EPSRC-funded ASSIST project for helping translators to address non-trivial translation problems (Babych et al. 2007). This functionality will be useful for linguistic terminological research or for research on contrastive grammatical theory for different languages.
- **Building lists of synonyms and related terms**: This can be done for individual words and for multiword expressions by discovering similar contexts in corpora (e.g., phrases or sentences in different texts that share certain features, like lexical items or syntactic structure). This functionality could be useful in language teaching tasks, e.g., for collecting material for grammar or vocabulary exercises. For this task we will integrate the free tool SemanticVectors (Widdows and Ferraro 2008).
- **Navigating a user-friendly interface to the search engine**: The various annotation and data analysis functions described above must be presented to the user consistently and simply. The interface will support multiple views on data in electronic corpora and accommodate the needs of a wide range of humanities research projects. For example, the use of corpora in language teaching projects requires the possibility of presenting complete texts to users and highlighting certain words or linguistic expressions in those texts. The interface will be developed on the basis of our Leeds CSAR tool (Sharoff 2006b), which will be extensively revised and updated to match the needs of humanities researchers.

These functions will be supported by documentation appropriate for researchers with a non-technical background.

The system will not be limited to these tasks. In the process of development we will work in close contact with humanities researchers in three target research projects, described in the *Exploitation and Collaboration* section. We will assess additional requirements for collection, analysis and presentation of researchers' data not envisaged at the current planning stage and implement new functions and relevant documentation specifically required for these projects. We expect that in the process of such collaboration new ideas for synergies between our computational tools and user content will emerge and we will endeavour to support them in our IntelliText system.

Research design

We have organised our work into the following activity streams, corresponding to work packages.

WP1. Tools for corpus collection and filtering: This work package will produce a set of tools for selectively downloading corpora from the web according to modifiable criteria specified by researchers, and cleaning up the collected texts. They will also permit the collection and maintenance of monitor corpora that are extended on a regular basis.

WP2. Integration of open-source tools for corpus annotation and alignment: In this work package we will bring together a set of free Open Source tools for different types of data annotation (part-of-speech, headwords, proper names), sentence and word alignment of translated texts. The main challenge is to assure compatibility between annotations, so that the different tools work together consistently.

WP3. Data search and analysis for electronic text corpora: This work package will extend existing tools for efficient and intelligent access to corpus data in such a way that they address the needs of researchers in the humanities. For example, the linguistic search engine will be extended to be able to work with sub-corpora (containing different genres or topics), and support additional levels of annotation, such as annotation of multiword expressions and linguistic constructions. It will integrate methods for generating collocation profiles, lists of synonyms and related terms, discovering similarities between texts segments and texts in corpora.

WP4. Flexible user interface and data presentation: In this work package we will develop an intuitive, user-friendly interface to the search engine and other functions integrated in the system. Researchers will be able to generate multiple views on data, targeting different uses of the data, e.g., for language teaching, discourse analysis, tone of voice and semantic prosody analysis. The interface will have a differentiated output, i.e. simpler, or basic, functions and then extended add-on functions for the more sophisticated users. In this way the interface will target a broader user base of researchers.

WP5. Documenting the software system: This work package will focus on providing a non-technical description of the tasks that can be performed by the system and will explore the principal ways in which researchers experiment with different data analysis methods (such as finding similar contexts or building time-lines for words or proper names) for different research projects in the humanities. In the work package we will create tutorials for various functions of the system, so that researchers can understand what they are and what they might use them for.

WP6. Liaison with target research projects and tasks: Here we will establish research networks with three target areas described above, in order to develop the system in such a way that it will be efficiently used and potentially enhance the impact of the target projects. We will explore together with humanities researchers synergies between the computational tools for intelligent access to their data and the specific content of their projects.

WP7. Dissemination of the results: This work package will, at the very start of the project, seek input from a panel of experts in a number of humanities disciplines in the form of prioritised desiderata for future applications. Before its conclusion, it will produce submissions to relevant conferences (as described in the justification of resources) and journals to disseminate the results to developers and potential user of the technology. The processing components and data will be released free of charge under General Public Licence, for both the research and the industrial communities.

The work plan is show in the following table:

	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10	m11	m12
1. Tools for collection												
2. Annotation and alignment												
3. Search and analysis												
4. User interface and presentation												
5. Documenting the software												
6. Liaison with target areas 6.1. Translation Studies 6.2. Language Teaching 6.3. Opinion / sentiment												
7. Dissemination												

Responsibilities of post-doctoral research fellows

Duties of Research Fellows on the project will include:

Post-doctoral Research Fellow responsible for computational aspects of the project

Focus: fundamental research on computational support for novel research methods in humanities, system planning, software design and implementation, maintaining the system's development cycle and designing a network for system's community support beyond the lifetime of the project

- Develop algorithms and new research methods for user-defined processing of corpora, including automatic annotation of relevant features and their and processing and visualisation for the users.
- Research and evaluate independently new linguistic annotation schemes and processing algorithms
- Define research objectives and specifications for software design and development at each stage of the project
- Setup an efficient communication network for project participants which will allow them to evaluate new features of the software system and to refine the requirements for each individual project task
- Communicate regularly with target researchers and research groups across the UK to establish a reliable feedback / implementation cycle for the project
- Design, plan and set up experiments to test performance of different modules, annotation schemes and algorithms
- Design and improve software prototypes and final software deliverables and documentation for target researchers of the project

- Organise meetings related to software development issues
- Make recommendations related to software development issues and software integration; participate in decisions regarding other aspects of the project.

Post-doctoral Research Fellow responsible for liaison with researchers in target areas

Focus: fundamental research on integration of computational corpus-based technologies into academic work in humanities; designing system specifications, creating system evaluation scenarios; liaison with target researchers, groups and projects

- Identify system requirements and to develop operational specifications of the computer system and its individual components based on discussions with target research groups
- Investigate independently principal objectives, quality criteria and usage scenarios for the system within the workflow of target scholars and research groups
- Devise methods for supporting innovative research projects in humanities and ensuring flexibility of the developed systems and its ability to work efficiently for a wider range of projects
- Proactively liaising with target researchers and groups, eliciting their needs for computational corpus-based support, ensuring synergies between novel computational methods in corpus linguistics and the core content of the target humanities projects.
- Set up an efficient communication network for the development/feedback cycle and to identify and reach out researchers who can benefit from using the system in future
- Establish testing and evaluation scenarios and criteria
- Make recommendations on models, concepts and features, which will be included into the system and to prioritise such aspects for system design and development
- Coordinate efficient collaboration between system development and its contribution to target projects, with a view to maximising their research impact.

Both Research Fellows will be expexted to:

- Independently prepare research papers and other project publications in their focus area within the project
- Update their own understanding of the needs of humanities researchers and the knowledge of advanced state-of-the-art methods in computational and corpus linguistics and actively apply them within the current project
- Evaluate the day-to-day progress of the project and suggest ways for meeting the objectives and addressing difficulties more efficiently
- Set up the directions for the project team in their respective areas (software design and creating efficient user workflow scenarios), in collaboration with PI and CoI
- Participate in preparation of further funding applications
- Contribute to feedback on research student dissertations related to the areas of the IntelliText project.

References

Babych, B., Sharoff, S., Hartley, A. & Mudraya, O. (2007) Assisting Translators in Indirect Lexical Transfer. Proceedings of ACL-07

Baroni, M., Chantree, F., Kilgarriff, A. & Sharoff, S. (2008) Cleaneval: a competition for cleaning web pages. Proceedings of the Sixth Language Resources and Evaluation Conference, LREC 2008, Marrakech

Bednarek, M. (2006). *Evaluation in media discourse: analysis of a newspaper corpus*. Continuum.

Butler, C. (2008). 'Basically speaking': a corpus-based analysis of three English adverbs and their formal equivalents in Spanish. In M. L. A. Gomez-Gonzalez, J. Lachlan Mackenzie & E. Gonzalez-Alvaraz (Eds.), *Current Trends in Contrastive Linguistics: Functional and Cognitive perspectives*. John Benjamins. pp. 147-176.

Christ, Oli. (1994) A modular and flexible architecture for an integrated corpus query system. COMPLEX'94. Budapest.

Ciobanu, D., Hartley, A. & Sharoff, S. (2006) Using Richly Annotated Trilingual Language Resources for Acquiring Reading Skills in a Foreign Language. *Proceedings LREC 2006*, Genoa. pp. 543-548.

Cunningham, H., Maynard, D, Bontcheva, K., & Tablan, V. (2002) GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia.

Corness, P. (2009) The treatment of the reporting verb said in Czech translations of English fiction: evidence of translation shift in InterCorp (to appear). Proceedings of InterCorp 2009.

Kurella, S., Hartley, A. & S. Sharoff. (2008). Rhetorical text structure in acquiring reading skills in L3. Proceedings of the 8th Teaching and Language Corpora Conference (TaLC 8), Lisbon. pp. 193-197.

Munday, Jeremy (in press) Evaluation and intervention in translation. In Mona Baker, Maeve Olohan and María Calzada Pérez (eds) *Advances in Translation Studies*. Manchester: St Jerome.

Munday, Jeremy (forthcoming a) Looming large: A cross-linguistic analysis of semantic prosodies in comparable reference corpora. In A. Kruger and K. Walmach (eds) *Corpus-Based Translation Studies*. Manchester: St Jerome

Munday, Jeremy (forthcoming b) *Evaluation in Translation: Critical points in translator decision-making*. Abingdon and New York: Routledge.

Och, F. J. & Ney, H. (2003) A Systematic Comparison of Various Statistical Alignment Models, Computational Linguistics, 29:1, pp. 19-51 March 2003.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of the International Conference on New Methods in Language Processing.

Sharoff, S. (2006a) Creating general-purpose corpora using automated search engine queries. In Baroni and Bernardini (Eds.). Wacky! working papers on the web as corpus. Bologna: GEDIT.

Sharoff, S. (2006b). A Uniform Interface to Large-Scale Linguistic Resources. Proceedings of LREC2006, Genoa.

Sharoff, S., Babych, B., Hartley, A. (2009) 'Irrefragable answers' using comparable corpora to retrieve translation equivalents. *Language Resources and Evaluation Journal*, 43: 1, pp. 15-25.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L. & Trón, V. (2005) Parallel corpora for medium density languages. Proceedings of RANLP2005, Borovets, Bulgaria. pp. 590-596.

Widdows, D. & Ferraro, K. (2008). Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application. *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.