Advanced corpus solutions for humanities researchers

James Wilson, Anthony Hartley, Serge Sharoff, Paul Stephenson

Centre for Translation Studies
University of Leeds, UK

Abstract This paper describes the design and implementation of an interface to corpora in 12 languages, stemming from the analysis of the needs of a diverse group of users: language teachers and language students, (non-computational) linguists, researchers in history and translation studies. We identified a set of requirements shared across the disciplines, as well as more specific requirements from the targeted user groups. The interface is designed to handle large-scale corpora of 20-500 million words.

Keywords: corpus interface, corpora, language teaching, history, translation studies

1 Introduction

This paper showcases the Intelligent Tools for Creating and Analysing Electronic Text Corpora for Humanities Research (hereafter, IntelliText) project that is being conducted by the Centre for Translation Studies (CTS) at the University of Leeds, UK. It describes how we have developed and extended our existing Leeds-based electronic corpora and tools [1] to meet the research demands of academics in various areas of the humanities.

After briefly outlining some general modifications to the interface and its documentation, we describe in detail how we have implemented the following functions:

- searches using metadata and statistics for metadata;
- automatic genre identification;
- advanced definition of shallow patterns;
- operations with frequency lists, including affix-based searches;
- classification of concordance lines according to their level of difficulty and appropriateness for language learners.

We have tested the tools with a range of corpora and languages, including representative web-derived corpora for Arabic, Chinese, English, French, German, Greek, Italian, Japanese, Polish, Portuguese, Russian and Spanish (Sharoff, 2006a), as well more specific collections, such as newswire or business corpora. In this paper, we present examples from Chinese, Japanese and Russian.

2 Outline of the project

IntelliText was conceived in order to allow humanities researchers, including those with little or no experience of working with electronic corpora, to make use of advanced methods of text collection and analysis. Rather than producing a new product we are developing our existing tools by enhancing the range of their functions and their usability. We are doing so by liaising with humanities researchers who represent several disciplines in order to cater for teaching and research needs across a spectrum of subjects. Our aim is to facilitate the use of corpora for a wide range of users, including students and non-specialists in corpus linguistics. Therefore, rather than determining on their behalf which new functions to add, we consultated corpus-active colleagues – IntelliText board members – from areas such as dialectology, language teaching, history and

[1] http://corpus.leeds.ac.uk/it/

linguistics in order to ascertain their needs at first hand. Given the limited lifetime of the IntelliText project, we adopted the principle that all modifications and enhancements should be applicable to more than one corpus language offered at CTS and should benefit users from more than discipline.

Much humanities research relies on or would benefit from analysis of electronic corpora. However, creating and annotating new corpora and the ability to uncover the information they contain are currently hampered by a lack of expertise in computing and database management and of experience in working with tools for data annotation and analysis. Most humanities researchers simply do not possess the required level of technical expertise; therefore, many projects, both small and large, miss opportunities for data analysis because of inadequate methodological or technological support. Similarly, even when a corpus already exists, the task of building or installing appropriate computational tools for analysing, intelligently searching and visualising the data still remains, for many humanities researchers, too challenging.

In the field of language learning, there is widespread agreement on the usefulness of authentic corpora, e.g., (Leech, 1997). While there is a considerable literature on using corpora in the teaching of English – for an overview see Section A10.8 in (McEnery et al., 2006) – corpus use has not yet entered the mainstream of English language teaching. The situation with other languages is considerably worse. For example, existing materials for teaching Chinese and, to a lesser extent, Russian often include grammatically correct but unnatural expressions. Progress in studying Chinese characters is often measured against “frequency lists”, while giving no evidence as to which corpora they are based on, how reliable they are, how genre-specific they are, etc.

In other fields of the humanities, many researchers are unaware how corpora could facilitate or benefit their teaching and/or research and they are unable to make use of corpus-based approaches that could increase the scope and impact of their projects. Moreover, computer scientists who design corpus-based tools are generally unfamiliar with the specific needs of humanities research; their tools are often difficult to adapt to a specific project or lack an intuitive interface and easy-to-follow documentation. Important potential synergies for the research of both parties have consequently been neglected. Therefore, as well as extending our tools in the ways already outlined, we are promoting them by demonstrating the benefits of corpus use in research and teaching at workshops and departmental seminars across the humanities.

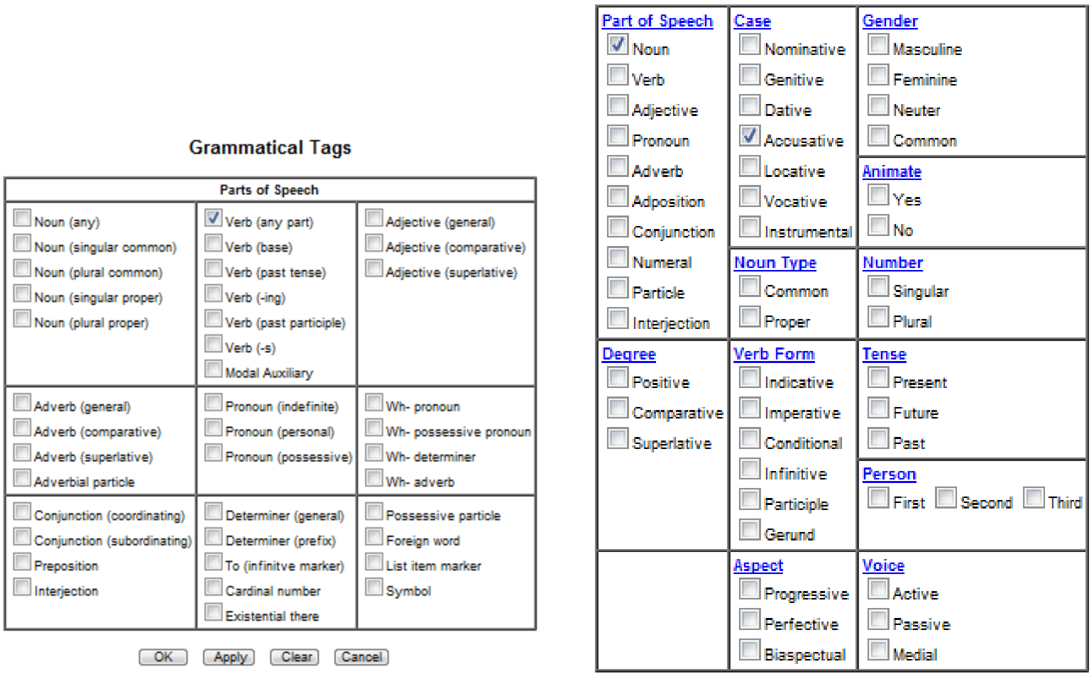
In sum, the IntelliText project sets out to:

- raise awareness of corpora and how they can benefit the research and/or teaching of academics in various areas of the humanities;
- promote the use of our corpora to academics with no previous experience of working with corpora and among students;
- improve the functionality of our corpora through liaison with academics across the humanities in order to understand and cater for their expressed requirements;
- make our corpora easier to use by improving the interface and documentation, providing illustrative searches for a variety of purposes, showing users in easy steps how to perform and interpret a corpus search, and offering a glossary of technical terms;
- enable our users to create and tag their own specialised corpora with the help of tools which are easy to use.

3 Updating of the interface and documentation

The tool at the core of the system is IMS Corpus Workbench (Christ, 1994), with a web-based query mechanism (Sharoff, 2006b). Word segmentation, lemmatisation and POS tagging is done using various language-specific tools, such as TreeTagger (Schmid, 1994) for European languages, MeCab[2] for Japanese, our own TreeTagger based tool for Chinese, etc.

[2] http://mecab.sourceforge.net/



Grammatical Tags

Parts of Speech		
☐ Noun (any) ☐ Noun (singular common) ☐ Noun (plural common) ☐ Noun (singular proper) ☐ Noun (plural proper)	☑ Verb (any part) ☐ Verb (base) ☐ Verb (past tense) ☐ Verb (-ing) ☐ Verb (past participle) ☐ Verb (-s) ☐ Modal Auxiliary	☐ Adjective (general) ☐ Adjective (comparative) ☐ Adjective (superlative)
☐ Adverb (general) ☐ Adverb (comparative) ☐ Adverb (superlative) ☐ Adverbial particle	☐ Pronoun (indefinite) ☐ Pronoun (personal) ☐ Pronoun (possessive)	☐ Wh- pronoun ☐ Wh- possessive pronoun ☐ Wh- determiner ☐ Wh- adverb
☐ Conjunction (coordinating) ☐ Conjunction (subordinating) ☐ Preposition ☐ Interjection	☐ Determiner (general) ☐ Determiner (prefix) ☐ To (infinitve marker) ☐ Cardinal number ☐ Existential there	☐ Possessive particle ☐ Foreign word ☐ List item marker ☐ Symbol

Part of Speech ☑ Noun ☐ Verb ☐ Adjective ☐ Pronoun ☐ Adverb ☐ Adposition ☐ Conjunction ☐ Numeral ☐ Particle ☐ Interjection	Case ☐ Nominative ☐ Genitive ☐ Dative ☑ Accusative ☐ Locative ☐ Vocative ☐ Instrumental Noun Type ☐ Common ☐ Proper	Gender ☐ Masculine ☐ Feminine ☐ Neuter ☐ Common Animate ☐ Yes ☐ No Number ☐ Singular ☐ Plural
Degree ☐ Positive ☐ Comparative ☐ Superlative	Verb Form ☐ Indicative ☐ Imperative ☐ Conditional ☐ Infinitive ☐ Participle ☐ Gerund	Tense ☐ Present ☐ Future ☐ Past Person ☐ First ☐ Second ☐ Third
	Aspect ☐ Progressive ☐ Perfective ☐ Biaspectual	Voice ☐ Active ☐ Passive ☐ Medial

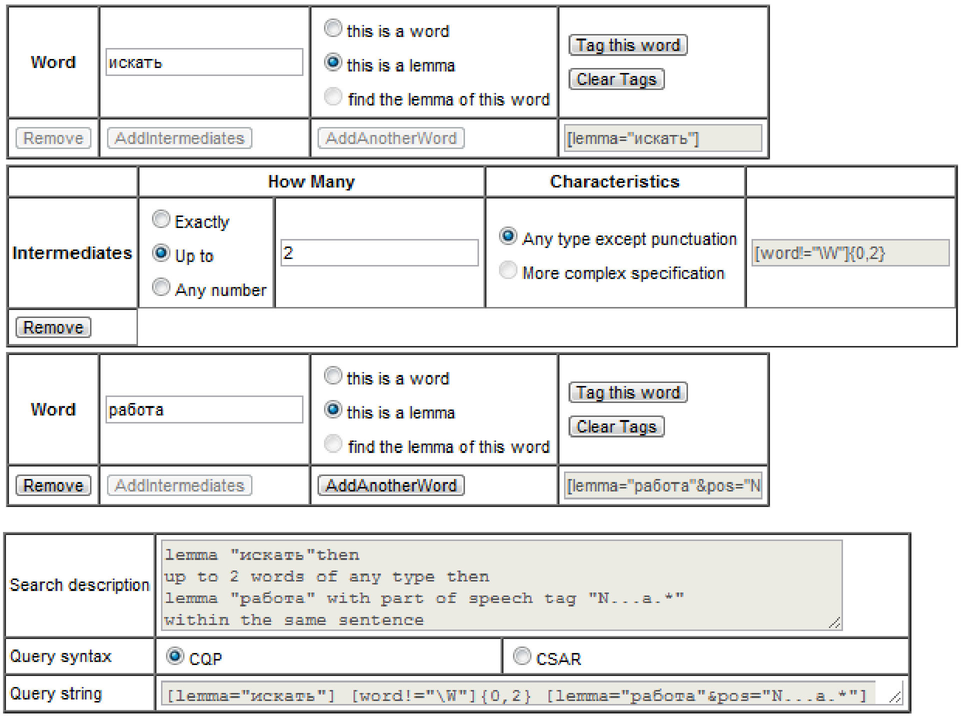
Figure 1: Selecting POS tags for English and Russian.

A qualitative survey of the IntelliText board members revealed that they considered the original interface counter-intuitive and they believed that it "put off" potential users. Many said that in their former state the Leeds-based corpora were inaccessible to students and to most non-specialists. For this reason, tutors used other, more intuitive interfaces on modules in which they introduced students to corpora and corpus-based research. Since tutors typically have only one lesson, or even just part of a lesson, to explain a particular tool or resource to their students, it is essential that the interface is intuitive and easy to use. Likewise, researchers do not want to spend weeks learning how to use a tool and they need an interface that is largely self-explanatory.

The survey also showed the need to completely revise the documentation. The instructions were not written in plain and simple terms, and some of string codes for the POS tags given in the documentation were wrong. This meant that first-time or inexperienced users could not necessarily perform all the tasks that they required – even having read the instructions. Moreover, the instructions were hard to find. In sum, a counter-intuitive interface combined with bad documentation made our corpora inaccessible to all but trained users.

Our task, therefore, was to create an intuitive interface supported by clear documentation, enabling the use of the tools by complete beginners. We have tried to make it possible for users to decide for themselves how much technical complexity they see by making only the most simple features visible by default and guiding users to more detail if they need it.

The new interface has several advantages. Perhaps the most significant advance is the ability to select POS tags by ticking check-boxes (Figure 1). This is much more user-friendly since it removes the need to know the relevant codes and so considerably reduces the chances of error in corpus searches. Other new functions include the option to show how many times a word or phrase occurs in the corpus and menus for creating frequency lists and performing multi-word searches. Entering full-fledged CQP queries for multi-word searches and complex patterns is usually beyond the skills of students or researchers in the humanities. Therefore, we have implemented the intuitive query builder interface shown in Figure 2. The actual query which will be submitted is also displayed, giving users the opportunity to learn the syntax if they wish.



Word	искать	○ this is a word ● this is a lemma ○ find the lemma of this word	Tag this word Clear Tags
Remove	AddIntermediates	AddAnotherWord	[lemma="искать"]

	How Many		Characteristics	
Intermediates	○ Exactly ● Up to ○ Any number	2	● Any type except punctuation ○ More complex specification	[word!="\W"]{0,2}
Remove				

Word	работа	○ this is a word ● this is a lemma ○ find the lemma of this word	Tag this word Clear Tags
Remove	AddIntermediates	AddAnotherWord	[lemma="работа"&pos="N

Search description	lemma "искать"then up to 2 words of any type then lemma "работа" with part of speech tag "N...a.*" within the same sentence	
Query syntax	● CQP	○ CSAR
Query string	[lemma="искать"] [word!="\W"]{0,2} [lemma="работа"&pos="N...a.*"]	

Figure 2: "Phrase search" function.

At the request of users we have introduced several new options for the KWIC concordance lines. For example, language teachers suggested that it should be possible to specify that the target word or phrase appears at the beginning or end of the extract as well as in the normal central position; this option is now available.

Another option suggested by users concerns varying the length of the context displayed for each occurrence of the key word in the concordance. Some researchers require a context much longer than the customary KWIC concordance, e.g., a few sentences or even a paragraph. These are users who expect search results similar to those from an Information Retrieval engine, which is nevertheless controlled by precise linguistic specifications. Their purpose may be, for example, to analyse the context in which a construction appears, or to show language students how a conjunction is used in the flow of authentic text. Lexicographers, in contrast, need to view a large number of lines simultaneously in order to detect patterns, thus a context of more than 5-7 words is unhelpful. Accordingly, we have included an option for selecting the number of concordance lines displayed on the concordance page.

The documentation has been updated in three ways. First, it has been made simpler and presupposes no prior knowledge of corpus linguistics; in fact, the documentation was written by researchers who are not computer scientists but who work in language teaching and linguistics and, as user advocates, can accurately gauge the level of detail that non-specialist users require. Second, a list of sample corpus searches is being provided for several corpus languages and in several subject areas in order to illustrate what types of search corpora are typically useful for. Third, a glossary of terms is being created provided, including expressions like *lemma*, *corpus*, *concordance line*, *collocation* and *POS tagging*. To facilitate use, we have included definitions of terms and POS tags in pop-up boxes which appear when the cursor is over them so that users do not have to navigate between pages.

4 Enhancing the functionality of the IntelliText tools

4.1 Automatic identification of genres and registers

A useful application of electronic corpora in language teaching is the ability to annotate collections of texts to show stylistic variation. Some traditional corpora, such as the BNC, contain fairly extensive annotation of their texts according to domains, audience types and genres (Lee, 2001). This information is normally not available for corpora collected from the Web. Even in traditional corpora, important genre or register distinctions may not be made at all or may be made in incompatible ways, rendering it impossible to show, for example, the difference between expressions of requests or suggestions in English and Japanese (e.g., -ましょ、-ませんか) in a given register.

We rely here on our current work on automatic genre classification (Sharoff et al., 2010), which can achieve reasonable accuracy provided that we have a manually annotated, topically diverse training sample of approximately 20-30 documents per genre or stylistic class on which to train the probabilistic classifier. The training documents within each genre need to cover a variety of topics. If this is not the case, even if a probabilistic classifier can achieve high accuracy in genre identification on the training set, it is more likely to be able discriminate between topics rather than genres in the rest of the corpus.

This facility has been incorporated to display elaborate stylistic tagging of Web corpora. For example, for a module in business Russian we applied formality tags so that business clichés can be extracted to help students differentiate between formal and neutral language use as well as to help them achieve a better understanding of register and to allow them to compile their own lists of generic phrases used in official documents. Similar tags can be applied to other languages, including Chinese and Japanese, to mark not only business clichés, but also more general formal versus informal oppositions, regional, colloquial, non-standard, archaic forms, and so on.

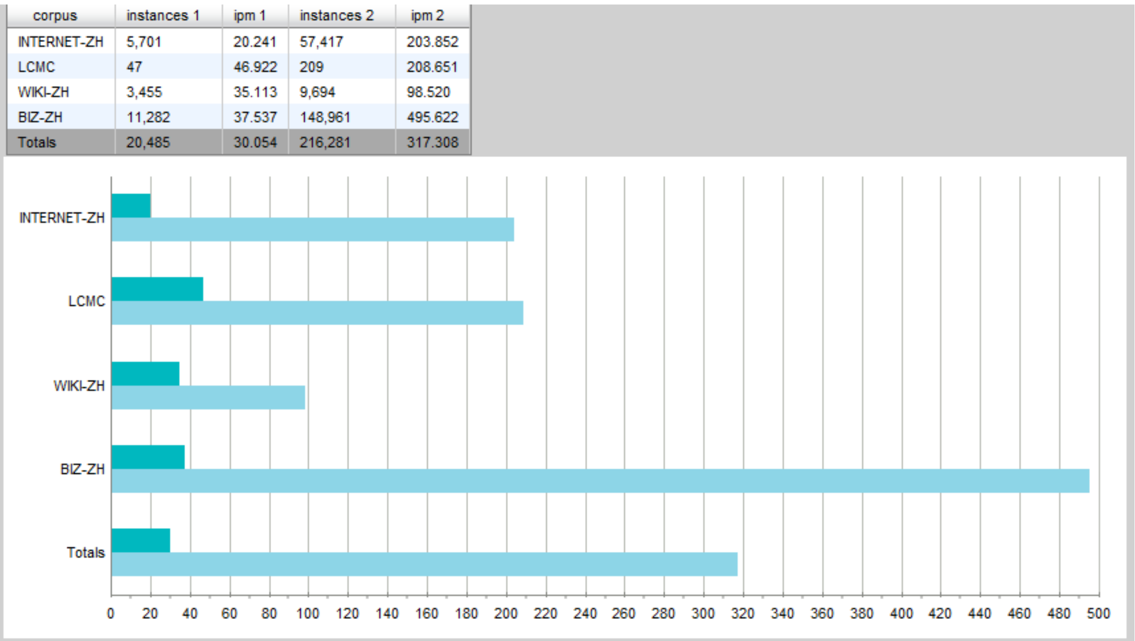
4.2 Exploiting metadata

Another function displays selected meta-data at the side of the concordance lines; users are able to select a variable such as "sex" and either M or F will appear next to the concordance line, enabling them to identify whether the passage was written by a man or a woman without having to manually inspect each example. For example, with respect to lexical differences, Lakoff argues that women use a wider range of adjectives of colour than men and that words like *mauve, lavender* and *ecru* are unmarked in what the author terms "women's language", while they are marked in "men's language" (Lakoff, 1975). It should be pointed out, however, that such claims are impressionistic; therefore, we can test their validity against corpora with specific metadata.

Similarly, monitoring uses of words in historical records implies the need to display the year or source of each retrieved context; therefore, we have made it possible to display concordance lines chronologically or using any other text provenance metadata. Finally, in a way similar to collocation lists which provide a sketch of the use of one word the context of other words, the interface makes it possible to sketch the use of a word with respect to metadata.

4.3 Frequency lists

There was previously no easy way of using our tools to generate frequency lists; researchers had to ask for them to be specially compiled. A "frequency list" option is now built into the interface, enabling any user to compile them quickly and easily. This feature is particularly useful for the teaching of Languages for Special Purposes (LSP) and for generating lists of the most common words from specialised sub-corpora. Vocabulary is especially important in LSP teaching, as in many domains a high level of language competence can be achieved only once an adequate discipline-specific vocabulary has been acquired. Thus, a prime application of such corpora for both language learning and translation is the extraction of terminology – single- and multiword terms – to create vocabulary lists of frequent words and collocations. Single-word terms are



corpus	instances 1	ipm 1	instances 2	ipm 2
INTERNET-ZH	5,701	20.241	57,417	203.852
LCMC	47	46.922	209	208.651
WIKI-ZH	3,455	35.113	9,694	98.520
BIZ-ZH	11,282	37.537	148,961	495.622
Totals	20,485	30.054	216,281	317.308

Figure 3: Comparison of frequencies for 情报 and 信息.

detected by loglikelihood scores for their frequencies against a reference corpus, while for multi-word terms we use an adaptation of the commonly-used Bootcat algorithm (Baroni and Bernardini, 2004). The corpora and derived word lists have provided a basis for teaching business Russian to British students.

4.4 Affix-based searches

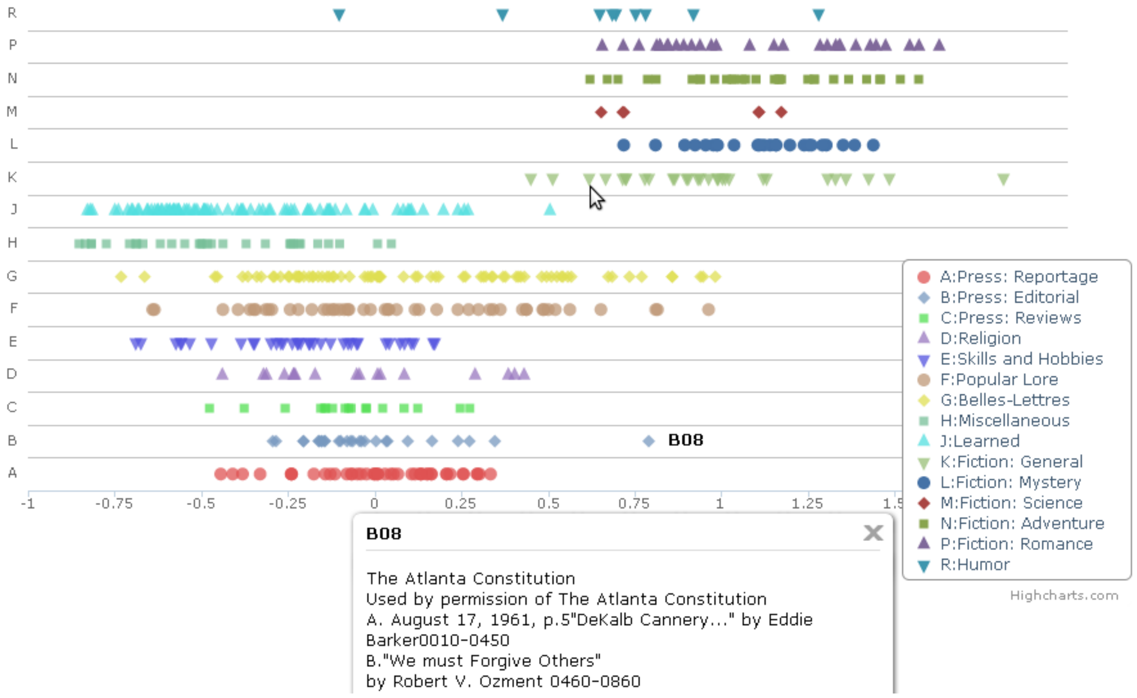
Initially the affix function was developed with learners of Russian in mind and was intended for verbal prefixes alone. However, it has a much broader application in teaching and research and is relevant to other languages, including English, German, Chinese and Japanese. The function now works also for suffix-based searches and is applicable to parts of speech other than verbs. Users may either enter a word and search for the prefixes or suffixes that are used with this word; or they may enter an affix and search for words that contain it. In both cases the output can be displayed either alphabetically or in terms of frequency, the most common prefixed form appearing at the top.

In the case of Chinese the learner needs to know in which multi-character words a given character is used. This functionality is routinely provided by Chinese dictionary software, but the frequency is not indicated. Yet frequency is crucial, otherwise the list may be populated with a large number of rare words containing characters unknown to the learner. For example, a prefix search for 面 in CEDICT[3] outputs 40 entries, some of which are infrequent and thus less useful for the language learner, such as 面庞 (‘facial features’, which in a representative corpus is outside of the top 20,000 most frequent lexical items). The frequency list can be also filtered by POS tags and linked to examples of use for each word. The frequency list can also be specific to a given corpus or subcorpus, e.g., business communication.

4.5 Statistical data

Several of our trial users have asked for more visualisations of word frequencies and other statistical information. To address these needs, we have added a number of simple features. It is now

[3] http://en.wikipedia.org/wiki/CEDICT



B08

The Atlanta Constitution
Used by permission of The Atlanta Constitution
A. August 17, 1961, p.5"DeKalb Cannery..." by Eddie Barker0010-0450
B."We must Forgive Others"
by Robert V. Ozment 0460-0860

Figure 4: Multivariate analysis along the 'narrativity' dimension in the Brown Corpus.

possible to see how many times the target word or phrase appears in total in the selected corpus (or corpora) and at the same time see a specified number of examples (Figure 3).

In addition, we have enabled users to perform more complex statistical tests on their data. In several disciplines, such as sociology and sociolinguistics, relative percentages and frequencies alone are of little practical value and researchers need to show whether these results are statistically significant. For example, it is possible to select two phrases and see whether there are significant differences in their usage. The same function can reveal, for example, the differences in use between the Japanese nominalising suffixes -ことが and -のが, whose use is largely complementary but partly overlapping.

This function is particularly useful for comparing competing translations or, in the case of language learning, for finding an appropriate collocate. Take, for example, the collocation *stiff competition*. A user can enter *competition* and one or more collocated adjectives – such as *stiff, rigid, solid* – and obtain the comparative statistical scores for these three collocations. Language learners will find this especially useful, given that conventional dictionaries are not ideal for collocations and literal translations from the source to target language often produce incorrect or odd collocations. Second, users can compare the use of words or phrases in different corpora. This allows researchers and students to compare, say, the occurrence and collocations of particular words in standard reference corpora to their occurrence in specialised corpora.

Another application of this function is the comparison of language varieties: users can easily identify (significant) differences between British and American English, European and Brazilian Portuguese, Peninsular and Mexican Spanish, and so on; moreover, as more regional corpora are being created, such statistical tests are particularly desirable, especially for sociolinguists. Third, we have made it possible to compare language use according to independent variables such as sex and age by allowing statistical comparisons of the meta-data, allowing researchers to show that men use a particular word or phrase *significantly* more than women, middle-aged speakers use word X *significantly* more than younger or older speakers, and so on.

Finally, our tools are capable of performing multivariate tests of a range of features against the set of documents in a corpus, or the features of a document against a corpus. This makes a data-driven analysis of genres (Biber, 1988) much more accessible to humanities researchers. In Figure 4 we give an example of a multivariate study of the Brown Corpus using the complete set of its POS tags rather than the features selected by Biber.One of the dimensions identified by the PCA transform can be related to Biber's narrativity: '0.249wPP+0.243wVVD+0.221wRP+0.220wVHD...', i.e., a higher load of personal pronouns and verbs in the past tense (the codes after 'w' refer to the Brown POS tags). The graph in Figure 4 shows the distribution of the Brown Corpus texts along this dimension according to the genre categories, e.g., fiction, as well as some memoirs and popular lore texts score higher on this dimension, whereas research papers (J) and government documents (H) have a much lower score. The interface can also show more information available for a selected node, e.g., for B08, the outlier in the category Press/Editorials. All of the statistical analyses that we have used, including the multivariate analysis, have been implemented with the freely available R environment (`http://www.r-project.org`) so that free distribution of the system is not compromised.

4.6 Filtering the concordance lines

In recent years, corpora have gained a strong foothold in learning and teaching. However, a major problem, highlighted by language learners and tutors alike, is that students often find it hard to understand the language written in the concordance lines; most corpora were not built with the language learner in mind and are made up of authentic texts of various genres and therefore no attention has gone into selecting texts of an "appropriate" level. Consequently, students at the beginner and (lower-)intermediate level have been highly restricted in what they can use corpora for.

Prior research in grading texts by their difficulty relied on assessing it from the viewpoint of native speakers of English, normally in the context of US schools or the US army (DuBay, 2004). In recent years there have been attempts to address the needs of learners of English as a foreign language, e.g., (Kotani et al., 2008; Kilgarriff et al., 2008; Heilman et al., 2008), but some of these studies relied on the availability of syntactic parsers or WordNet, and none of them addressed the needs of learners of other languages.

In prior research (Sharoff et al., 2008) we established parameters for assessing the difficulty of texts and individual sentences in several languages (English, Chinese, German and Russian) by comparing the parameters associated with texts judged to be more or less difficult by language tutors for these languages. These parameters were detected by using a Principal Component Analysis (PCA) transform of a large set of features, which were easy to compute with minimal resources for each language. The parameters correspond to the two most significant dimensions:

- the lexical complexity estimated by the coverage of a text or a line by a frequency band, e.g., `top2000` refers to the amount of text covered by the 2000 most frequent words (lemmas) in a given language;
- the syntactic complexity estimated by counting the average number of prepositions, lexical verbs and conjunctions per sentence and the average sentence length in words (`asl`).

In the interface, the language tutor can use two sliders to control the difficulty thresholds, which filter the concordance lines according to different learner levels.

5 Conclusion

Thanks to extensive changes to the interface and the creation of clear and comprehensive user-manuals, together with other supporting materials, non-specialists can quickly and easily perform basic corpus searches as well as carrying out more complex tasks that before only computer experts and trained corpus linguists could do. Our interface enables users to generate frequency lists, carry out complex multi-word searches and perform several statistical tests; moreover, users are now able to collect, annotate and analyse their own collections of texts, benefiting from the range of functions that our tools offer. In this respect, we have taken corpora from the exclusive domain of specialists into the mainstream, thus offering academics working in the humanities (and beyond) possibilities for facilitating and enhancing their teaching and research.

A user-oriented approach in the design and development of our tools has produced a powerful interface with an array of new and improved functions. Liaising with academics from various departments and schools within the humanities, we have been able to tailor our tools to several target groups and cater for corpus-based research and teaching needs in several disciplines and for several languages. In particular, we have extended the functionality of our tools in the areas of statistics and tagging. We now possess an elaborate tagging system that can be used for genre classification, register tagging and filtering concordance lines in terms of their level of difficulty and usefulness for the language learner. In addition, we have added advanced statistical possibilities that allow users to test for significant differences in the use of individual words and phrases, collocations, competing translations, etc., according to independent variables and across corpora; they can also perform multivariate analyses.

To date, researchers and teachers have typically had to use several interfaces to perform all the functions that they need. IntelliText has made a major contribution to corpus-based research and teaching in that we have brought together in a single package many of the features that academics working in the humanities require. Furthermore, in so doing, we have added functions not available in other interfaces. These include: affix-based searching, ranking concordance lines by their difficulty, automatic genre classification and multivariate analysis. Finally, since the IntelliText tools will be made available as an open-source resource, they can be enhanced or modified according to changing trends and new demands in research and teaching, to cater for disciplines beyond the humanities, and with regard to new technological developments.

Acknowledgements

This work has been supported by AHRC DEDEFI grant AH/H037306/1.

References

Baroni, M. and Bernardini, S. (2004). Bootcat: Bootstrapping corpora and terms from the web. In *Proc. of LREC2004*, Lisbon.

Biber, D. (1988). *Variations Across Speech and Writing*. Cambridge University Press.

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

DuBay, W. (2004). The principles of readability. Technical report, Impact Information.

Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *Proc. the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79, Columbus, Ohio. ACL.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proc. of EURALEX'08*.

Kotani, K., Yoshimi, T., Kutsumi, T., Sata, I., and Isahara, H. (2008). EFL learner reading time model for evaluating reading proficiency. In *Proc. of CICLING*, Haifa.

Lakoff, R. (1975). *Language and Woman's Place*. Harper and Row, London and New York.

Lee, D. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, 5(3):37–72.

Leech, G. (1997). Teaching and language corpora: A convergence. In Wichmann, A., Fligelstone, S., McEnery, A. M., and Knowles, G., editors, *Teaching and Language Corpora*, pages 1–23. Longman, London.

McEnery, T., Xiao, R., and Tono, Y. (2006). *Corpus-based language studies: an advanced resource book.* Taylor & Francis.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester.

Sharoff, S. (2006a). Creating general-purpose corpora using automated search engine queries. In Baroni, M. and Bernardini, S., editors, *WaCky! Working papers on the Web as Corpus*. Gedit, Bologna. http://wackybook.sslmit.unibo.it.

Sharoff, S. (2006b). A uniform interface to large-scale linguistic resources. In *Proc. of the Fifth Language Resources and Evaluation Conference, LREC 2006*, pages 539–542, Genoa.

Sharoff, S., Kurella, S., and Hartley, A. (2008). Seeking needles in the Web haystack: finding texts suitable for language learners. In *Proc. of Teaching and Language Corpora Conference, TaLC 2008*, Lisbon.

Sharoff, S., Wu, Z., and Markert, K. (2010). The Web library of Babel: evaluating genre collections. In *Proc. of the Seventh Language Resources and Evaluation Conference, LREC 2010*, Malta.