SCORE: A Multimodal Corpus Database
of Education Discourse in Singapore Schools

Huaqing HONG[1]
National Institute of Education,
Nanyang Technological University

A large-scale effort to sample classroom interactions, teaching materials and students' assignments in Singapore schools has been undertaken by the Centre for Research in Pedagogy and Practice (CRPP), National Institute of Education, Singapore. In line with the goals of a range of research projects related, a computer corpus approach to the exploration of such a large collection of education discourse is applied. It is thus critical to compile a multimodal, multi-level annotated corpus database for the sake of various qualitative and quantitative analyses of classroom interactions, curriculum materials and students' oral, print and electronic artefacts. In this paper, we first briefly introduce the demand and significance of compiling such a database valuable for any large-scale empirical investigation of education discourse. Next, the compilation work itself poses challenges at various stages, including data handling, corpus structure/scheme design, discourse annotation and pattern modelling, query tools development, and data filtering and retrieval. Whenever possible, we propose the solutions to a list of problems encountered in processing the speech and textual data for the database. Finally, we discuss in what way the database can be queried with a variety of query tools, such as a web-based query interface, a client-server web-service tool, and third-party corpus tools.

1. Introduction

Hargreaves (1996) made a trenchant criticism of educational research in the UK by saying that "in education there is simply not enough evidence on the effects and effectiveness of what teachers do in classrooms to provide an evidence-based corpus of knowledge." he stated that educational researchers tend to undertake small scale investigations that produce inconclusive and contestable findings of little practical relevance. He suggested that the core reason for this is the way in which educational research is commissioned and funded, which is insufficiently focused on the classroom, and does not involve teachers in a sufficiently meaningful way. Similarly, in reviewing the education research in the US, Biber et al (2004) pointed out that "to date there have been few large-scale empirical investigations of academic registers, and virtually no such investigations of spoken academic registers." Given this lack of large quantity of data as support, it has been impossible to generalize the educational research results on the basis of traditional pen-and-paper approach or observation-based approach (cf. Cazden, 1988; Markee, 2000; Christie, 2002). Specifically in the context of classroom research, we have lacked the empirically defined classroom interaction models drawn on a large sample of data.

To bridge such a gap, the Centre for Research in Pedagogy and Practice (CRPP), National Institute of Education, Singapore, has undertaken a large-scale effort to sample classroom interactions, teaching materials and students' assignments in Singapore schools since 2003. This effort eventually made it possible to compile a computerised corpus

---

[1] Contact detail: 1 Nanyang Walk, National Insittute of Education, Singapore 637616. Phone: (+65)67903368
Fax: (+65)63164787   Email: hqhong@gmail.com or hqhong@nie.edu.sg

database of education discourse, the Singapore Corpus of Research in Education (SCORE). With the availability of such a database, various goals of a range of research projects can be achieved with a computer corpus approach to the exploration of education discourse. We believe that a multimodal, multi-level annotated corpus database can benefit education researchers interested in both qualitative and quantitative analyses of classroom interactions, curriculum materials and students' oral, print and electronic artefacts. Therefore, this paper presents the SCORE project, its significance, design, development and application. Concerns and problems with regard to the development of such an education discourse database are also reported.

2. Objectives and Educational Applications

With the availability of a computerized corpus of education discourse, especially the collection of a large quantity of classroom interactions, researchers can conduct large-scale empirical investigations of classroom register. In brief, the main goal of the SCORE project is to provide a large general purpose resource for the education researchers from home and abroad. In specific, we attempt to achieve the following objectives in this corpus database project:

- to build a multimodal multilevel annotated classroom discourse corpus database;
- to develop a sophisticated query engine for education researchers to do corpus-based empirical studies on classroom discourse;
- to develop a bundle of automated discourse annotation and query tools;
- to provide statistical models of Singaporean school classroom interactions;
- to present and publish academic papers in international journals and conferences.

In line with these objectives, the deliverables will include a speech subcorpus, a lexical subcorpus, and several multilevel annotated subcorpora at different stages. Eventually all those subcorpora will be indexed and incorporated into a well-designed corpus database, which will be provided with sophisticated query tools for both online and offline queries. In sum, the deliverables of this project will consist of three kinds: 1) annotated corpora (time-aligned speech corpus, lexical corpus, parts-of-speech tagged corpus, semantic features tagged corpus, speech-act tagged corpus, systemic functional linguistic features tagged corpus, and spoken discourse annotated corpus); 2) computer tools (conversion tools, semi-/automatic discourse annotation tools, database query engine, web-based query tools, statistical analyzer of classroom discourse, etc.); and 3) technical reports, corpus/database manuals and academic publications.

With these deliverables, the potential educational applications of this project can be carried out from the following several perspectives:

A multimodal classroom corpus: Digital audio/video recordings of the classes make it possible to conduct in-depth classroom research on knowledge acquisition and pedagogy. Along with the accompanying transcriptions and classroom codings, the digital recordings will be available for corpus analysis for many years. Materials drawn from these annotated classroom recordings can be used in professional development and material development. Using a streaming program, researchers and teacher trainers can extract and view media files along with associated coding and transcription data for research and professional

development activities. With the query tool, users can search the database for clips of media illustrating particular points of knowledge acquisition or pedagogy. Linguists, who are interested in acoustic analysis of language use in the classroom, can also make use of such a multimodal classroom corpus database (see Gordon et al, 2004, for a similar study).

A multilingual and multidiscipline corpus: The corpus sampling of SCORE project covers several curriculum subjects (Language, Literature, Science, Social Studies, History, Geography and Mathematics) mediated in four official languages (English, Chinese, Malay and Tamil) in Singaporean primary and secondary schools. A corpus-based investigation of linguistic variation within or across class sessions and disciplines can find the associated patterns of language use and classroom practice. Such corpus-inspired research areas likely include:

- Language use in/for the classroom;
- Corpus-based contrastive analysis (across languages and subjects);
- Corpus-based genre/register analysis (in comparison with other domains);
- Learner language (e.g. code-switching, code-mixing, error analysis and more);
- Learning and teaching strategies;
- Insights for designing effective teaching materials and classroom activities;
- Implications for effective use of teaching aids (IT, slides, OHT, etc.)

Empirical model of classroom practice: There are many practices that are believed to be particularly useful in classrooms, but whose effectiveness has not been empirically demonstrated. Many of these can be investigated in the searchable database of classroom interactions. The areas of particular interest are: techniques that facilitate various kinds of student-student and student-teacher interactions; activities that facilitate literacy development and knowledge acquisition; and the learning gains associated with various types of instruction.

The resource of teacher training aids: With the abundance of sample classroom interactions collected and annotated, teachers would gain new perspectives about task design and input that would best lead to a good practice, along with an understanding of how learners may grasp one level of meaning rather than others. They may also benefit from the effective practices demonstrated in the database, and eventually develop a better awareness of what good practices should be. Classroom transcripts can be used to develop awareness of, and promote practice in, the language used for various categories of teachers' verbal behaviours. The use of lesson transcripts in teacher training can also develop teachers' skills in asking questions in the classroom and their understanding of the pedagogical role teachers' questions play in the foreign language classroom (Cullen, 2001; Trappes-Lomax & Ferguson, 2002; Han & Cong, 2003).

Critical classroom discourse analysis: Conversation analysts tend to reject simulated or "set-up" interactions from experimental situations (Burns, 2001). With the availability of authentic materials collected from real classroom interactions, researchers interested in conversation analysis can get access to meticulously recorded, carefully transcribed and linguistically annotated samples of natural data. The application of this project can help researchers and practitioners gain an understanding of how to develop reflective, evidence-based practice to improve student learning outcomes. It can connect

practitioners, policy-makers and researchers by providing models of interaction and feedback loops, so that each group can enhance the work of the others, as well as work toward continuous improvement.

3. Compiling The Corpus Database

Basically, the design of the corpus database of education discourse consists of three phases of tasks, namely, 1) data collection and manipulation; 2) feature selection and corpus annotation; and 3) database building and query. This is roughly presented in the development diagram above (see the next page).

3.1. Data Collection and Manipulation

We will process the SCORE corpus data from different sources at different stages and integrate them into a well-designed relational database system. The data for the corpus are mainly from four sources, which include audio/video recordings of classroom lessons, teaching materials (textbooks, PowerPoint slides, OHT transparency, etc.), students' artefacts and classroom coders' coding sheets. The machine-readable transcripts of the 500 hours classroom lessons are one of the most important data source at the initial stage, and the corpus is deemed to grow with more data collected by the group.

To prepare the raw data for the corpus annotation and database query, we developed a series of Perl Scripts/Programs that actually convert the raw transcriptions into several formats, including machine-readable format and human-reader-friendly format. In this way, the Perl program can automatically correct some commonly-occurred conventional transcription errors, and it can also generate a list of non-convention tokens/words, which consists of misspelled words (e.g. *didnt* for *didn't*), localized words and terms (e.g. *Boon Lay*, *Jurong*, *Cher* (localized short form for addressing "teacher")), and quasi-lexical vocalisations (e.g. *hmm, lah, aha, shssh*, etc.). The unknown-word list, for one thing, helps us standardize the transcription convention of spoken discourse; and for the other, it can minimize the error rate in the corpus annotation followed. In this way, the data will as far as possible be cleaned up to minimize or eliminate errors.

In addition, to protect privacy, an anonymiser was created to programmatically anonymise the personal names, place names, business names and so forth, which otherwise in any way would then be liable under the laws of the country in question when the corpus is released to the public. However, this doesn't mean that those proper nouns are removed from the corpus, which is actually flagged in some way.

To assist the annotation process, we have to develop a conversion tool to convert raw files before loading into MMAX (Multi-Modal Annotation in XML), an XML-based linguistic annotation editing tool (Müller & Strube, 2003). This conversion tool can be done by examining the structural conventions of MMAX file format and relating the various attributes in the file to the corresponding points in the source file. By implementing algorithms that translate the vital points in the source files to attributes in the MMAX file format, we are able to convert files from any source to the MMAX format.

To enhance the integrity of the data, it is important to ensure that there is no data loss during the data conversion process. This can be achieved by extracting all possible

information that we have in the first place with the audio/video recordings, and then placing it in the appropriate portions of the target data format. In cases where there is no way of incorporating the extra information in the target data format, we will generate a file containing all extraneous data. This file will serve to be an index for the program that imports data into the database; the program will then use this index to put the data into the database with proper referencing to the original data.

Finally, as presented in the diagram below, a bundle of MMAX output files are supposed to be consolidated and incorporated into the final corpus database. A java program reads the output files from MMAX and imports them into the database. To make the database accessible to a large audience, the final output files are encoded in accordance with the standard XCES XML format (see *XCES Corpus Encoding Standard for XML* for detail).

3.2. Feature Selection, Annotation Tools and Methods

To make use of such a resource of education discourse, a researcher will expect detailed yet reliable information from the corpus, and thus the annotation of the corpus data with linguistic and paralinguistic features will be the critical part of the corpus database compilation. Linguistic annotation covers any descriptive or analytic notations applied to raw corpus data (Bird, 2001). The added notations to the media or textual basic data may include transcriptions of all sorts, from phonetic features to discourse structure. The focus is on labelling linguistic information which has been widely used for constructing annotated linguistic databases, and on the formats commonly adopted by such tools and databases.

To annotate the corpus data, several issues have to take into account: what kind of linguistic features should be annotated? In what way these features can be annotated? Can the annotation work be done automatically? Should the features/levels be annotated one after another or simultaneously? How much time and labour does the annotation work demand? Drawn on *The Cambridge Grammar of the English Language* (Huddlestone & Pullum, 2002) and *An Introduction to Functional Grammar* (Halliday & Matthiessen, 2004), these annotated linguistic features (see the selected features in the table below) range from morphosyntatic to discourse, including part-of-speech tags, semantic categories, thematic/modal types, nominalization types, transitivity types, clause types, speech-act types, interclausal relations, adjacency pairs, initial-response-feedback (IRF), topically related sets (TRS), phase/episode changes, etc.

As a general purpose education research corpus, it is ideally to annotate as many as possible features for different research goals. However, the linguistic and paralinguistic features are interwoven from different perspectives of linguistic theories or schools. Many of those features cannot be automatically annotated with handy computer tools, yet semi-automatic processing or editing tools are helpful. In regard to the features to annotate, we have surveyed the existing tools in the literature and selected a few suitable tools (see Table below) to assist our annotation (see Sim, Kazi & Hong, 2005, a survey of discourse annotation tools).

SCORE CORPUS DATABASE DEVELOPMENT DIAGRAM
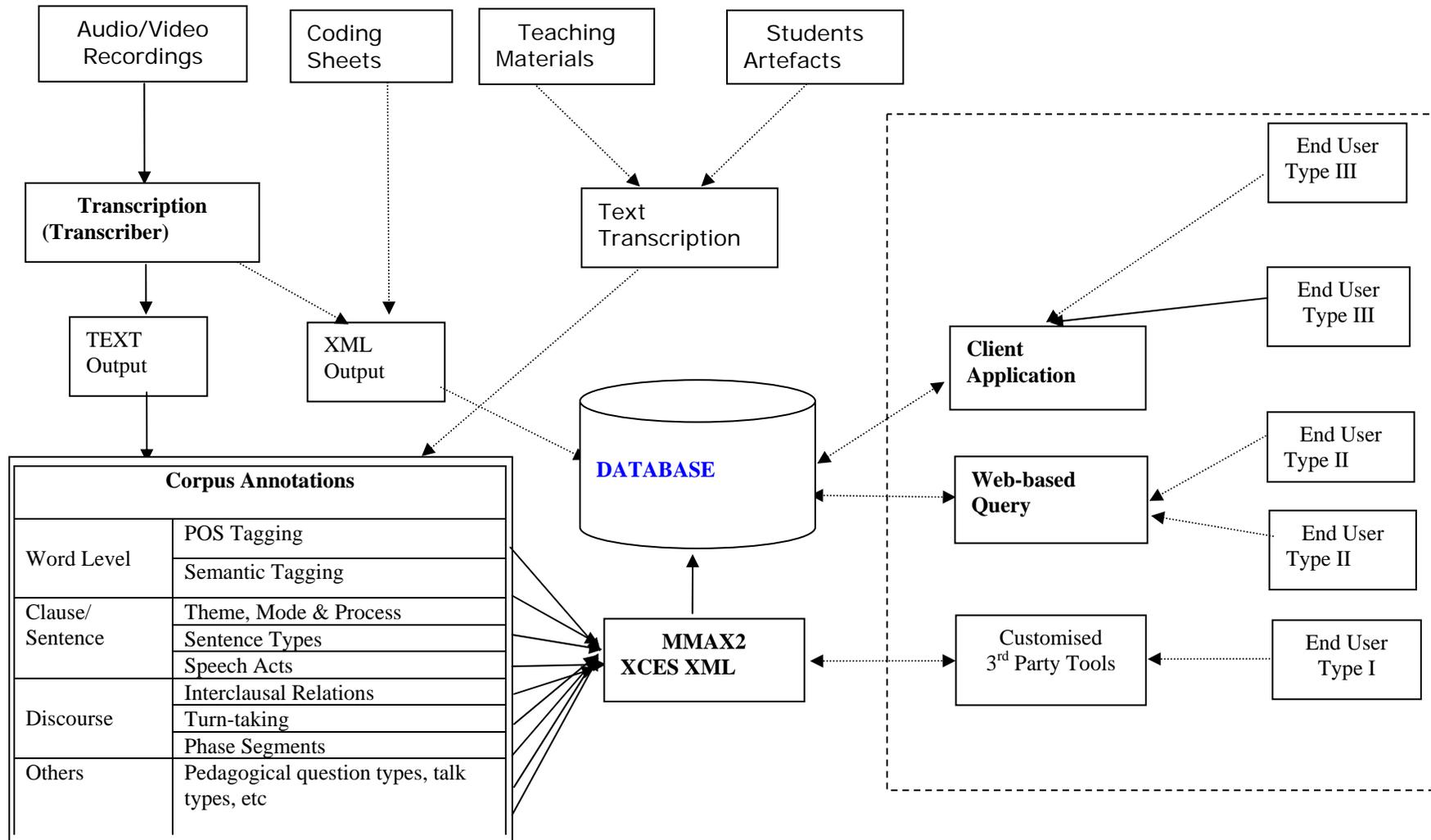
LEVELS	TAGGING	EXPLANATION	TOOLS
Token Word Phrase	POS Tagging	Automatic tagging, with manual post-check unknown words	CLAWS Tagger, (USAS) (English) ICTCLAS (Chinese) NSP
	Semantic Tagging		
	N-gram Lexical Patterns		
Clause Sentence	Theme, Mode & Process	Systemic Functional Grammar (SFG) features	MMAX2
	Sentence Types, Nominalizations	SFG and Speech-Act Annotation Scheme	MMAX2
	Speech Acts		MMAX2
Discourse	Interclausal Relations	SFG	MMAX2
	Turn-taking, interactions	Dialogue analysis	MMAX2
	Phase/episode segments, IRF, TRS, adjacency pairs	Can be done in transcribing	MMAX2
Others	Utterance types, text reading, teaching or learning strategies, code-mixing and switching, etc	Nature of input materials, knowledge structure, instructional means, etc.	Transcriber / MMAX2

Table: list of the selected features and annotation tools used in the SCORE project

The design of the corpus has been negotiated between different interest groups, regarding such questions as what annotation schemes to adopt, and what standards or what guidelines to adhere to. It will be clear that, in order to be able to complete such a large corpus within the time constraint, many transcriptions and annotations have to be done in parallel rather than sequentially. A factor which complicates matters is the fact that different researchers require different levels of annotations. In practice, we have to complete one layer of annotation for a certain file before it is made available for another layer, though it is still subject to change as errors may be reported from other annotation layers. In this way, working in parallel annotations certainly has its own merits. Moreover, this method can make it possible to achieve a better inter-annotator agreement rate between the group of annotators, and it can also make us to fine-tune the annotation schemes accordingly and revise them before too late.

3.3. Database Building and Query

In the final phase we will develop our database in order to facilitate researchers to query the corpus data. In this way, a list of query tools and applications will be developed. At the initial phase, all the corpus data with corresponding multi-level annotations are stored in a nested directory structure either in plain text or in XML format. Although each item of data can be queried with different third-party tools or custom-made programs, this poses a steep learning curve for education researchers and other potential users. So, it is of importance to build a user friendly yet robust query interface to the SCORE database.

3.3.1. Initial Query

To meet the request from some researchers' interests in exploitation of the data before the database is ready for any extensive use, we need provide some simple query tools or customize a few third-party tools for the researchers to query the corpus. The list of third-party corpus query tools includes Wordsmith tools, Xaira, XQuery, Concordancer, and others. These query tools are commonly available and have their own advantages and

disadvantages (Sim, Kazi & Hong, 2005). These initial query tools and their sample results can facilitate our understanding of different researchers' wants and needs from the corpus. Moreover, we can also test the database design as well as the proposed sophisticated query tools.

It might seem imprudent to build these initial query tools instead of building the sophisticated query tools right away. However, this is necessary as a launching platform to build the sophisticated query tools. The underlying query code base can be shared with the final sophisticated query engine. Therefore, there is no duplication of work.

These tools serve as an initial point for researchers to perform simple query on a pre-filtered corpus. At this point, the annotated data, along with the corpus is in the XML format, which ensures data portability. Handling of the data requires only simple technical knowledge of computers and the raw data can be viewed with a multitude of programs. These initial query tools will also have their uses after the database and the sophisticated query tools are built. They will serve as a means of offline query tools for researchers who are unable to connect to the network for some reason.

3.3.2. The Database

After investigating the state-of-the-art technologies in XML and relational database systems, we found that queries to relational database are much faster than queries to XML files (Wittenburg et al, 2005). Furthermore, a well designed relational database will definitely help us manage SCORE data better. There are other major considerations taken into account when deciding on taking a database approach. Here are the list of concerns in constructing the database.

Database updates: With a database system, updating of the data in the database is a lot less labour intensive and error prone. This is due to the underlying mechanism of the database which performs checks to validate and verify the change in the database. There is a continual need to update the database as we receive new data from time to time, or make corrections to the existing database. These operations are a continual process that builds up the corpus, and data integrity is of paramount importance. We cannot afford to have slip-ups in this area.

Speed of query: Databases are built for purposes of storing and retrieving data. Flat files cannot hope to compete with dedicated database systems in this aspect. Some of the queries that we will be implementing such as frequency distributions and modelling of the classroom require a large amount of data throughput to generate results. A fast query will reduce the amount of time waiting for the system to return data, especially on a multiple user database.

Reliability: Database systems are known to be more reliable than flat files. Flat files can be accidentally modified, moved or deleted without permissions, causing a failure in the system. Such failures are usually accompanied by a tedious process of rectifying the cause of failure, resulting in downtime for researchers. To modify data in the database, one will need to have proper permissions to perform the necessary operations on the database. Internal checks by the database will also ensure that the data is not fouled by users. Data rollback can be performed on the database when something goes wrong.

Multiple user support: Databases are designed to be accessed by many users, especially over a networked environment. The database management system will take care of the user interactions on the database such as updating the database and retrieving data. It ensures that all users will have the most updated copy of the database without the problems that are usually associated with multiple user environments such as file locking, or cross updating of data.

Single copy of data: With a single copy of the data, it is much easier to keep track of the proper versioning of the data. Backing up and doing maintenance on the database becomes much easier with a single copy of the database.

Control of the data: With a database system, we can control how much data to release to the users of the database. Compared to the flat file structure, this promises much more security of data, ensuring that only those with permissions can access the data. We will also have a record of users who accessed the database, therefore having a safety net to control how data is released.

Support transactions: Database support transactions which keep track of the state of user queries. This feature is important for users who have multiple queries, and each query is dependent on the results of the previous queries. This feature will aid the researcher who is researching in a very specific area of the classroom discourse.
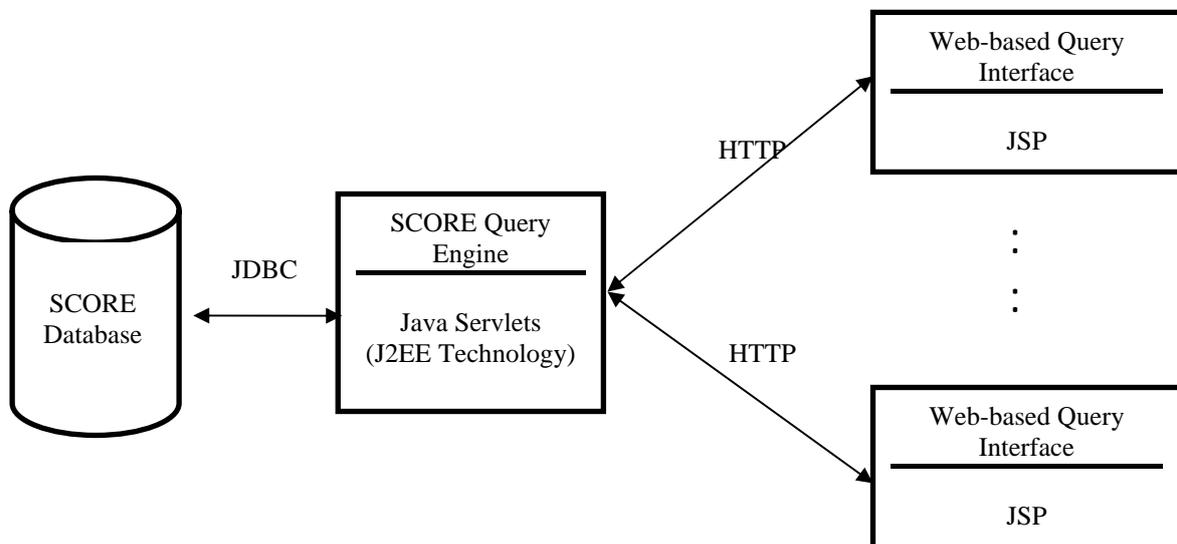
3.3.3. Web-based Query

The above figure depicts the architecture of the web-based query engine we adopt. A web-based query interface to SCORE database is an ideal approach to make this corpus available to education researchers and other potential users, because with such an approach:
- the database resides on the server side, and remains transparent to the users;
- users do not need to download and install any client program, and just need to login into the designated webpage and query the database.

We will basically provide two types of web-based query for the users: a lazy query with a wide range of selectable fixed parameters; and an advanced query with wildcards or regular expressions.

A web-based query is intuitive as it presents all options to the user in an easy to read format. Compared to a text based system, a Graphical User Interface (GUI) has a smoother learning curve for users. As we cannot expect our target users to have solid query knowledge, a GUI based query interface will empower them to retrieve the data in an intuitive, simple and efficient way.

In the web-based query, relevant data are grouped together, allowing users to easily put down their mental picture of the data into a concrete form. Valid options for each field are presented to the user in a form of drop down menu of options, eliminating the possibility of users generating an illegal or ill-formed query. Therefore, a GUI-based query will reduce frustrations faced by users of computer systems.

A web-based query will also aid us in controlling the release of data to users of this system. Users have to log in to this system in order to gain access to the data. We can then decide, based on the user profile, how much data the user is allowed to access. In this way, the data is presented to the user on a "need-to" basis, allowing users with proper permissions to access the data.

All that a user needs to access the web-based query tool will be a web browser (Internet Explorer, Netscape, Firefox, Opera, etc.) and an internet connection. The operating costs are almost negligible since users of the system need not purchase any additional software. This is certainly a plus factor for users of the system. Additionally, any user with an internet access can use the query tool, allowing them to do investigation wherever comfortable.

3.3.4. Advanced Applications

The advanced application of web-services interface allows users the privilege to manipulate the database. Technology savvy users can then build software using the web-services interface to retrieve relevant data, and process the data accordingly. The SCORE corpus is not necessarily used only by human users, other software tools, such as NIEtro (Diederich, Pederson & Pederson, 2005), may also query SCORE database. We will develop web-services interface to our SCORE query engine, so the legitimate client software can query SCORE using this Web-Services interface. Using such web-services can extend the usage of the corpus far beyond that envisioned by us. It would be an important tool to unleash the limitless potential of the corpus.

4. Concluding Remarks

In this paper we reported our on-going project of compiling a multimodal corpus database of education discourse in Singapore schools. We have outlined the significance and design of the corpus, the annotation schemes and methods, and the computer tools regarding data manipulation, linguistic annotation and database query. The aims set in this SCORE project are by all means quite ambitious, yet the construction of such an education discourse database can fill the gap in the educational research, which has been criticised for lacking an

evidence-based corpus of knowledge to justify the effects and effectiveness of teaching practices in classroom. It should be noted that building a corpus of such would be extremely demanding, given the fact that it is highly labour-intensive and time-consuming in terms of data preparation, linguistic annotation and database generation. In addition, to cater for various research purposes from different interest groups from home and abroad, we attempt to annotate as many as possible features, which makes the task more complicated and challenging. However, it is worthwhile for the effort to work toward a systematic, empirical description of classroom practice with various corpus-based qualitative and quantitative analyses. Therefore, as the largest computerised education discourse database in the region, and probably in the world, we believe that our effort and experience can also benefit a larger research community from various perspectives.

References

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). Representing Language in the University: Analysis of the TOEFL 2000 spoken and written academic language corpus. (ETS TOEFL Monograph Services MS-25). Princeton, NJ: Educational Testing Service.

Bird, S. (2001). Linguistic Annotation. Available on-line from http://www.ldc.upenn.edu/annotation/ (accessed March 20, 2005).

Burns, A. 2001. Analysing spoken discourse: Implications for TESOL. In Analysing English in a Global Context: A reader. Burns, A.(ed.). London: Routledge.

Cazden, C. (1988). Classroom Discourse: The language of teaching and learning. Portsmouth, NH: Heinemann.

Christie, F. (2002). Classroom Discourse Analysis: A functional perspective. London: Continuum International Publishing Group Ltd.

Cullen, R. (2001). The use of lesson transcripts for developing teachers' classroom language. System, 29: 27-43.

Diederich, J., Pedersen, M., & Pedersen C. (2005). NIEtro: Shallow topic detection and tracking for the analysis of classroom interactions. Proceedings of the Conference of Redesigning Pedagogy: Research, Policy, Practice. Singapore, May 30 - June 1, 2005.

Gordon, E., Campbell, L., Hay, J., Maclagan M., Sudbury, A., and Trudgill, P. (2004). New Zealand English: Its Origins and Evolution. Cambridge: Cambridge University Press.

Halliday, M.A.K. & Christian M.I.M. Matthiessen. (2004). An Introduction to Functional Grammar (3rd revised edition). London: Hodder Arnold.

Han, X., and Cong, T. H. (2003). Teachers' language use in Hong Kong Mandarin classroom. In New Perspectives on Teaching Chinese. Cong, T. H. (ed.). Beijing: Beijing University Press.

Hargreaves, D. (1996). Teaching as a Research Based Profession: Possibilities and Prospects. London: Teacher Training Agency.

Huddlestone, R. & Pullum, G. (2002). The Cambridge Grammar of the English Language, Cambridge, Cambridge University Press.

Markee, N. (2000). Conversation Analysis. London: Lawrence Earlbaum Associates.

Müller, C. & Strube, M. (2003). Multi-level annotation in MMAX. Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue, Sapporo, Japan, 4-5 July 2003, 98-107.

Sim, T. J., Kazi, S. A., & Hong, H. (2005). A survey of the transcription, annotation, and query tools for the development of a classroom discourse corpus. Proceedings of the Conference of Redesigning Pedagogy: Research, Policy, Practice. Singapore, May 30 - June 1, 2005.

Trappes-Lomax, H., and Ferguson, G. (2002). Language in Language Teacher Education. Amsterdam and Philadelphia: John Benjamins.

Wittenburg, P., Broeder, D., Piepenbrock, R., & Veer, K. (2005). Databases for Linguistic Purposes: a case study of being always too early and too late. Available on-line from http://emeld.org/workshop/2004/Wittenburg-paper.html (Accessed March 9, 2005).

XCES Corpus Encoding Standard for XML. Department of Computer Science, Vassar College, USA. Available on-line from http://www.xml-ces.org/ (Accessed Feb 20, 2005).