A Survey of Transcription, Annotation and Query Tools for Development of a Classroom Discourse Corpus

TZE JAN SIM SABBIR AHMED KAZI HUAQING HONG

Centre for Research in Pedagogy and Practice
National Institute of Education, Nanyang Technological University

ABSTRACT

A multilevel annotated classroom discourse corpus is being developed by the Centre for Research in Pedagogy and Practice (CRPP). This corpus will allow a range of related research projects in CRPP to perform various qualitative and quantitative analyses. The development of this corpus has several considerations, among them, the need to select suitable tools is critical. As there are many ready-made tools available, selecting the right tool can cut down on a substantial amount of development time. In this paper, we shall evaluate a list of tools for audio/video transcription, corpus annotation (e.g. parts of speech, semantic, system functional grammar) and database query. The list of tools we surveyed mainly consists of three types: transcription tools (e.g. Transcriber, NITE, Anvil, TASX, etc), annotation tools (both manual and automated, e.g. MMAX, Systemics, Wmatrix, YamCha, etc) and query tools (e.g. TEC, Concordancer, Wordsmith, etc). They will be benchmarked based on a number of key requirements of our project. These requirements include interoperability with other tools, support of task-specific annotation schemes, customization flexibility, and general user friendliness. Finally, we will describe how output from the selected tools will fit together, and the additional work required to achieving seamless integration. We believe that such a survey can also shed light on the research community interested in processing spoken discourse.

INTRODUCTION

The Center for Research in Pedagogy and Practice (CRPP) has undertaken a project to build a multilevel annotated classroom discourse corpus, SCORE (the Singapore Corpus of Research in Education). Building such a corpus is a time consuming and laborious task. From transcription to annotation to query and analysis, efforts in every stage of corpus building can be greatly reduced by using well designed tools which serve their niche purposes.

SCORE is a large scale effort requiring collaboration and coordination between different groups of people consisting of transcribers, annotators, software engineers and linguists. Specialized tools with data interoperability are essential for effective teamwork. Data interoperability allows data to be transferred easily between tools to be processed.

This paper presents a survey and evaluation of selected tools that perform transcription, annotation and query functions. We developed a set of criteria for comparison of each group of tools based on SCORE's requirements. We believe, however that these criteria are relevant not only to SCORE, but also to the many projects that is similar to SCORE in nature dealing with multilevel annotated discourse. However, we recognize that some tools, while performing very well on their own, might not have the capability to export data for other tools to use. Therefore, the corpus team has decided to build some data format conversion tools to bridge the tools together.

TRANSCRIPTION TOOLS EVALUATION

In this section, we present our criteria for evaluating the transcription tools which we believe to be of relevance for similar transcription efforts.

We shall list out the criteria we applied in choosing a transcription tool for our purposes. Note that our criteria do not test for highly detailed features of the tool, but rather we will examine the tools with the criteria that will suit our needs for a transcription tool.

Except for Transcriber, the rest of the tools are multipurpose tools which support annotation and query functions. However, we will only be looking at the transcription capabilities of the software listed.

Table of transcription tools criteria evaluation

Tool Criteria	**Transcriber**	**Praat**	**AGTK**
Modality	Audio	Audio	Audio
Supported audio input formats	WAV, AU, AIFF, MP3, CSL, SD, SMP, and NIST/Sphere	AIFF, AIFC, WAV, AU, NIST	WAV, AU, AIFF
Unicode support	Yes		Yes
Cost/License issues	Free software under GNU General Public License	Free software under GNU General Public License	OSI-approved Common Public License
Media control	Millisecond control with bar and buttons	Millisecond control with bar	Millisecond control with buttons and bar
Interface	Layman	Layman	Layman
Coding scheme support	Tags can be changed	None	Requires programming skills
Type of coding	Time-stamped	Time-stamped	Time-stamped
Import/export	Export to .typ, .stm or .trs (XML structure)	Export to other sound formats	Interface to WaveSurfer
Metadata	Limited	None	Can be programmed

Tool Criteria	**Anvil**	**TASX**	**NITE**
Modality	Audio and Video	Video Audio	Video Audio
Supported audio input formats	AIFF, AU and WAV	AIFF, AU and WAV	MP3, WMA, VOB, AIFF, AU and WAV
Unicode support		Yes	
Cost/License issues	Free for research purpose		Freeware
Media control	Millisecond control with buttons and bar	Seconds control with buttons and bar	Millisecond with buttons
Interface	XML skills required for adding coding scheme	Layman	Layman
Coding scheme support	Requires XML skills	None	Interface provided
Type of coding	Time-stamped, structure	Time-stamped	Time-stamped
Import/export	Import from Praat, Xwaves, export to SPSS	Agtk, exmaralda, praat, anvil syncwriter, transcriber	Export to XML
Metadata	Limited	limited	Free form metadata can be entered

We need to select a transcription tool based on SCORE's requirements, we shall discuss the requirements in this section. XML output from the tool is preferred, XML allows data to be portable, thus allowing data to be shared easily among other programs (Jean et al., 2002). XML is a well documented data format that can be easily understood by many software packages. Therefore it is easier to share data with other tools that we are going to use in our project. The requirement for a time aligned transcript is also important, this will allow us to extract segments of audio based on the transcripts when the query tool for the corpus is up. The tool should also support overlapping speech. In the area of classroom discourse, overlapping speeches are common and therefore, we would want a

transcription tool that allows overlapping speech to be represented. The tool should also allow transcripts to be entered in Unicode, as we will be dealing with multilingual transcripts, and in Singapore's context, a single transcript might sometime consist of several languages being spoken in a single utterance. Metadata should be a part of the transcripts, as we would want the speaker's demographic/background information to be entered as part of the transcript. Having phase-change information would be essential when we are looking out for linguistic features for a particular phase in the discourse. User defined event markup allows us to markup events as required, so that we can retrieve these markers from database for research later.

Based on our requirements, we have chosen to use Transcriber to transcribe audio. Transcriber meets most, if not all of our requirements for a transcription tool.

ANNOTATION TOOLS EVALUATION

Annotation tools, both manual and automated are evaluated in this section. We present our results which we believe to be of relevance for similar annotation efforts.

A well designed annotation tool which adapts itself to the various annotation tasks at hand can ease the job of the annotator.

Some of the tools listed have additional functionalities, however, we will be looking only at the text annotation functions of the software.

Table of manual annotation tools criteria evaluation

Tools Criterion	**TASX**	**EXMaRaLDA**	**MMAX**	**PALinkA**
Data				
Preprocessing	Optional	Optional	Obligatory	Obligatory
Unicode	Yes	Yes	Yes	Yes
Markables	Start-end	Start-end	Start-end	Inclusion
Atomic features	Yes	Yes	Yes	Yes
Relation between markables	No	No	Set	Pointer
Dominance relations	No	No	No	Bracketing
Metadata	Yes – as annotation	Yes – as annotation	Yes – as annotation	Yes – as annotation
Interoperability				
Import/Export of annotation scheme	No	No	Yes - XML	Yes - Text
Converter	Yes	Yes	No	No
Plugins	Yes	No	No	No
Specifying Annotation Schemes				
Annotation levels	No	No	Yes	No
Annotation Tagsets	No	No	Yes - Structured	Yes
Specification	No	No	External	External
Annotation process				
Automatic Annotation	To an extent	No	No	To an extent
Selection based	No	No	Yes	Yes
Visualization				
Scope of annotated work	All	All	Focus	Focus
Style	Text	Text	Choice menu	XML
Additional highlighting	Yes – User defined	Colouring font type and size	Yes – User defined	Colouring brackets
Reference units of additional highlighting	Value	Feature	Feature/Value	Feature
User adaptation	Tier hiding	Tier hiding	Yes	Yes
User Definition	Yes	No	Yes	Yes

Tools Criterion	**Systemics**	**AGTK**	**DialogueTool**
Data			
Preprocessing	Optional	Obligatory	
Unicode	Yes	Yes	
Markables	Inclusion	Inclusion	
Atomic features	Yes	No	No
Relation between markables	No	No	No
Dominance relations	No	No	No
Metadata	Yes	Yes	
Interoperability			
Import/Export of annotation scheme	Yes - Text	No	Yes
Converter	No	No	No
Plugins	No	No	No
Specifying Annotation Schemes			
Annotation levels	No	No	No
Annotation Tagsets	Yes - Structured	No	Yes
Specification	Internal and External	No	External
Annotation process			
Automatic Annotation	No	No	No
Selection based	Yes	No	Yes
Visualization			
Scope of annotated work	Focus	All	Utterance level only
Style	Text	Text	Text
Additional highlighting	Colouring	Colouring	
Reference units of additional highlighting	Value	Feature	
User adaptation	No	Yes	
User Definition	No	Yes	

SCORE annotation requirements

Our requirements for an annotation tool for use in SCORE entails many different aspects of usability and functionality of an annotation tool. The need for a flexible display and interface mechanism with customizable schemes and tagsets is especially important. It is obvious to those who have worked with multiple codings that a certain display will work well for one task, but useless for another task. The ease of configuration allows rapid customization for display or usage, allowing annotators to customize the tool to their own needs and preferences. Customizable schemes and tagsets allows annotators to choose features and schemes specific to our domain and needs. Annotated data visualizations allows the annotator to see which parts of transcripts are annotated at a glance, easing his job of annotation without the need to check on the transcript to see if he has missed out on any annotation. Unicode support is paramount as there are multilingual discourses to be annotated. Support of multi level annotation allows for different levels of linguistic features to be annotated. Using of a single tool to annotate different levels of linguistic features allows us to streamline our data processing techniques of the annotated data. By using XML as a data format, it will allow data portability (Jean et al., 2002). Researchers will be able to communicate data to those who are using other software packages, or use this data as a better software as it comes along. We have chosen XML as the data structure as it is universal and well documented with many software packages available to extract data from the file. In SCORE, the possibility for group work becomes important when there are multiple annotators

working on the same piece of transcript for different linguistic features. It is preferable that the tool allows group work without the risk of a member of the team corrupting the data of another teammate's work.

Based on our requirements, we have chosen to use MMAX as our annotation tool. It has a good developer support, and the schemes and visualization are fully customizable. Most importantly, it uses separated XML files to store annotated data, which allows annotators to work on a same piece of data without corrupting the work of other annotators.

Automated Taggers

Automated tools can reduce the workload of annotators drastically, although automated taggers are not entirely error-proof, the same can be said for human annotators injecting human errors. Our aim is to reduce the workload of human annotators.

However, there are few reliable automated taggers around, and even fewer had been used for real world annotation. Automated taggers still have a long way to go before they can replace human annotators. We will still evaluate the taggers that we have found and determine their usefulness in SCORE.

Table of automated tagger tools criteria evaluation

Taggers / Criterion	**TNT**	**YamCha**	**Perl Lingua**	**Wmatrix**
Functionality	PoS Tagger	PoS Tagger, Named Entity Recognition, NP chunking	PoS Tagger	PoS and Semantic Tagger
Method of tagging	Statistical	Kernel Based (SVM), PKE (Polynomial Kernel Expanded), PKI (Polynominal Kernel Inverted)	Probability, bigram (two-word) Hidden Markov Model	Probablistic
Flexibility of language and tags	Can train with any language and tags	Can train with any language and tags	Uses Penn Treebank Tagset	English only, tagset defined
No of Tags	any	any	44	60-160 depending on tagset
Tagging Performance	It is typically between 30,000 and 60,000 tokens per second on a Pentium 500, running Linux.	1-2 sec / sentence, faster speeds can be achieved by using PKI and PKE. (PKE faster)		15 mins on 25,000 words on high end machine
Features		Can redefine feature sets (window-size), parsing-direction (forward/backward) and algorithms of multi-class problem (pair wise/one vs rest) Can perform partial chunking		Uses CLAWS and USAS, to produce POS and Semantic annotation
Output	txt, one token per line	txt, one token per line		XML
Programming lang		C/C++ library	Perl	
OS	linux	linux	Any with perl	SUN OS4x or UNIX
Accuracy (claimed)	94.5-96.7	93-94		96-97

SCORE automated tagger requirements

For automated annotation tools, we would want something that is easy to use without much customization, the accuracy of tagging and the reputation of the tagger. From this requirements, Wmatrix (consists of CLAWS and USAS) emerged a clear winner. Wmatrix has been used in tagging a couple of corpora, including BNC. The XML output of Wmatrix also helped as it allows us to use the annotated data easily.

QUERY THE CORPUS

Developing a multi-level linguistic annotated speech corpus, an essential component of research and development in human language technology (HLT), is a costly and resource-consumable task. In order for this effort to pay off, flexible access to this speech corpus database is a must and is possible through some custom-made or generic query tools. Structure of such a query tool is very much subjective to the data model used to develop the annotated speech corpus. Any descriptive or analytic notations embedded to the raw language data or speech transcript is considered as "linguistic annotation". The basic speech data is usually situated in a form of sequential time functions (e.g. audio recordings) while the descriptive or analytic notations are embedded hierarchically which can cover from phonetic features to discourse structures including phonetic segmentation and labeling, POS and semantic tagging, syntactic bracketing, 'name-entity' identification, co-reference annotation, prosodic phrasing, intonation and gesture information and so on. So, the complexity of an annotated speech corpus's data model depends on the levels of annotations that have been implemented. Based on the data-model adopted, usually a multi-level annotated speech corpus has its own special-purpose query tool. However, efforts have been made to develop general purpose data model, e.g. Annotation Graph, to create generic query tool with some promising results. In this section we will discuss about the query criteria for annotated speech corpus and survey the existing query tools.

Query Criteria

Queries to a multi-level annotated speech corpus fall into two broad categories: statistical queries and logical expression queries. The statistical queries include word-frequency, entropy, etc., whereas the logical expression queries are usually based on natural query language to perform various types of searches including pattern search, regular expressions search, feature search, word search, extract by metadata (class structure, speaker info), proximity search, concordance, collocations, etc. Some queries have well-formed format such as word-frequency, concordance, etc. These kinds of queries can be done through pre-defined query statements. Some queries are open, such as <*find all the utterances that contain the word 'OK'*>. These kinds of queries require a query language coupled with a query engine and a display or storage for query results.

Query Tools

There are handfuls of tools publicly available for predefined queries such as word-frequency, word-list, concordance, etc. These include WordSmith (Scott, 1998), Concordance (Watt, 2005), TEC (Ronaldo, 2005), etc. These tools perform the predefined queries based on well established algorithms. There are also a wide range of query tools available for performing logical expression queries for speech corpus. Here is a concise survey of these tools:

Emu: Emu (Cassidy & Harrington, 2001) query tool is to search Emu Speech Database. This speech database is annotated into a set of levels and levels are organised into hierarchies. Emu supports intersecting hierarchies and the query results usually processed with an external statistical package.

MATE: MATE (McKelvie et al., 2000) has a query tool (Q4M) to search spoken dialogue corpora which is annotated with MATE data model. Like Emu, MATE also supports intersecting hierarchies. Query results are stored in XML documents.

Annotation Graph: Annotation Graph (AG) (Bird & Liberman, 2001) data model aims to develop a general purpose scheme for linguistic annotation. An AG consists of a set of nodes, representing time points, and a set of level arcs, representing tokens. Queries in the AG query language are made up of a set of path expressions which describe paths through the annotation graph. AG query tool supports a wide range of annotation data model such as TIMIT (Garofolo et al. 1986), Partitur (Schiel et al. 1998), CHILDES (MacWhinney 1995), LACITO (Jacobson et al. 2001), LDC Telephone Speech, Switchboard (Godfrey et al. 1992), Emu, MATE, etc.

NXT: NXT (Heid, et al. 2004) query tool is developed to search NITE corpus, but it also supports other annotation data model (both time-aligned and hierarchical) by converting them into NITE model.

MMAX Discourse API (Müller & Strube, 2002): This is an approach towards the development of reusable software components for discourse processing tasks. This API (implemented in Java programming language) set can perform query to the corpus annotated with MMAX annotation tools. Our investigation on this API set shows that this is still in the developmental stage.

Besides the above mentioned speech corpus query tools, there are a number of text corpus query tools also publicly available. Among the most important ones are:

CorpusSearch (2005): A query tool for searching Penn-Helsinki Corpus of Middle English, which is based on Penn Treebank data model. This tool can be used for any corpus which is annotated in the Penn Treebank style.

ICECUP III (Wallis, Aarts & Nelson, 1999): This query tool is designed to query ICE-GB, British component of the International Corpus of English, which is based on tree data structure.

NetGraph (Mírovský & Ondruška , 2002): This is a corpus-workbench for Prague Dependency Treebank.

TIGERsearch (Brants et al., 2002): Developed to query a German newspaper Treebank. It also supports other data model such as NEGRA and Penn Treebank.

TGrep2 (Rohde, 2005): This is a Unix Grep style query tool developed for querying corpora annotated in Penn Treebank format.

VIQTORYA (Steiner & Kallmeyer, 2002): A query tool for Tübingen Treebank.

In SCORE, we have our own data model which is based on the output data from Transcriber and MMAX. Since we have developed our own coding scheme in MMAX and incorporate speaker's demographic data from the Transcriber's output, there is no readily available query tool that can serve our query purpose. So, we decided to develop our own query tool for the SCORE corpus, which is a client-server application using Java server programming framework. On the client-side it has a query-builder interface along with a set of predefined queries. Corpus query engine is implemented in a series of Java servlets.

CONCLUDING REMARKS

In this paper, we have presented selected transcription, annotation and query tools that can be applied for building a multilevel annotated classroom discourse corpus. On the list of requirements, we have developed a list of criteria for these tools. After inspecting the results of this evaluation, we have chosen our desired tools based on the criteria that it fulfils. The tools that we have chosen performs their individual tasks well, and can be integrated into the larger part of corpus building. However, we still need to work on a data conversion program in order to merge the tools nicely together. Our selections of the tools are of course, based on our requirements of SCORE, potential users are encouraged to evaluate the tools further with their own specific requirements after comparison based on the information provided in this paper.

REFERENCES

Bigbee, T. and Loehr, D. and Harper, L. (2001) Emerging Requirements for Multi-Modal Annotation and Analysis Tools, In: Proceedings of Eurospeech, pages 1533-1536.

Bird, S. and Liberman, M. (2001) A formal framework for linguistic annotation. Speech Communication, 33:23–60

Brants, S., Dipper, S., Hansen, S., Lezius, W. and Smith, G. (2002) The TIGER Treebank. In Proceedings of the Workshop on Treebanks and Linguistic Theories Sozopol.

Carletta, J.; McKelvie, D.; Isard A. (2002). Supporting linguistic annotation using XML and stylesheets. In Readings in Corpus Linguistic, G. Sampson; D. McCarthy, Continuum International.

Cassidy, S. and Harrington, J. (2001) Multi-level annotation in the Emu speech database management system Speech Communication, 33, 61-77, January, 2001.

CorpusSearch. (2005) http://corpussearch.sourceforge.net/index.html, accessed on April 20, 2005.

Dipper, S., Gotze, M., and Stede, M. (2004) Simple annotation tools for complex annotation tasks: an evaluation. In Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora, pages 54–62, Lisbon.

Dybkjær, L. and Bernsen, N. O. (2004) Towards General-Purpose Annotation Tools–How far are we today?. In Proceedings of the LREC.

Garg, S., Martinovski, B., Robinson, S.,Stephan, J., Tetreault, J., and Traum, D.R. (2004) Evaluation of transcription and annotation tools for a multi-modal,multi-party dialogue corpus. In Proceedings of the LREC.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., and Dahlgren, N. L. (1986) The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM. NIST.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992) Switchboard: A telephone speech corpus for research and development. In Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing, volume I, pages 517–20

Heid, U., Voormann, H., Milde, J-T, Gut, U., Erk, K., and Pado, S. (2004) Querying both time-aligned and hierarchical corpora with NXT Search. In Fourth Language Resources and Evaluation Conference, Lisbon, Portugal, May.

Jacobson, M., Michailovsky, B., and Lowe, J. B. (2001) Linguistic documents synchronizing sound and text. Speech Communication 33, 2001, p. 79-96.

MacWhinney, B. (1995) The CHILDES Project: Tools for Analyzing Talk. Mahwah, NJ: Lawrence Erlbaum., second edition.

McKelvie, D., Isard, A., Mengel, A., Møller, M. B., Grosse, M. and Klein, M. (2000): The Mate Workbench - an annotation tool for XML coded speech corpora. In Speech Communication, 33(1-2), pp97-112.

Mírovský, J. and Ondruška , R. (2002) NetGraph System: Searching through the Prague Dependency Treebank . In Prague Bulletin of Mathematical Linguistics , pp. 101--104. MFF UK .

Müller, C. and Strube, M. **(**2002**)** An API for Discourse-level Access to XML-encoded Corpora. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands, Spain, 29-31 May, pp. 26-30.

Rohde, D. (2005) TGrep2 User Manual (http://tedlab.mit.edu/~dr/Tgrep2/tgrep2.pdf - accessed on April 20, 2005)

Ronaldo. (2005) TEC: An Open-Source Concordancing Tool: http://ronaldo.cs.tcd.ie/ - accessed on April 20, 2005

Schiel, F., Burger, S., Geumann, A., and Weilhammer, K. (1998) The Partitur format at BAS. In Proceedings of the First International Conference on Language Resources and Evaluation.

Scott, M. (1998) WordSmith Tools Version 3.Oxford University Press, Oxford, England.

Stefanie Dipper, Michael Gotze, and Manfred Stede. Simple annotation tools for complex annotation tasks: an evaluation. In Proceedings of the LREC Workshop on XML-based Richly Annotated Corpora, pages 54–62, Lisbon, 2004.

Steiner, I. and Kallmeyer, L. (2002) VIQTORYA -- A Visual Query Tool for Syntactically Annotated Corpora. In Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Gran Canaria, May 2002.

Wallis, S. A., Aarts, B., & Nelson, G. (1999). Parsing in reverse – Exploring ICE-GB with Fuzzy Tree Fragments and ICECUP. In J.M. Kirk (ed.), Corpora Galore. Papers from the 19th International Conference on English Language Research on Computerised Corpora, ICAME-98 (pp. 335-344). Amsterdam: Rodopi.

Watt, R. J. C. (2005) Concordance package. Useful hints, tips, and links: http://www.concordancesoftware.co.uk - accessed on April 20, 2005-04-20.