The Blackfoot Language Database

This paper describes the Blackfoot Language Database (BLD¹), an online linguistic fieldwork database for researchers studying the Blackfoot language.

Fieldwork linguists amass huge quantities of primary linguistic data at considerable cost. In many cases, and especially when highly endangered languages are involved, there is great demand for these data both from researchers and from the communities of the languages under study. To a significant extent, advances in theoretical linguistics, language documentation, pedagogical materials development and language revitalization efforts are dependent on efficient use of primary linguistic data.

The thesis of this paper is that primary linguistic data can be used more efficiently and that a tool like the BLD can help. In the following sections, I explore in greater depth the ways in which primary linguistic data are inefficiently used and the benefits of a structured and easily accessible store of high quality linguistic data such as the BLD.

INEFFICIENT USE OF PRIMARY LINGUISTIC DATA.

I see three principal challenges to the efficient use of primary linguistic data: (i) finding the appropriate data, (ii) sharing the data amongst researchers, educators and community members and (iii) increasing the likelihood that data elicited for a particular purpose can be used for other purposes.

After having spent many hours transcribing and recording utterances and linguistic judgments, the linguistic researcher is faced with the sometimes daunting task of sifting through their data and finding the appropriate examples to support their theoretical argument or to illustrate a grammatical feature. When the mass of data are stored as hand-written notes or textual files (be they plain text, Word documents or PDFs), it is not uncommon, in my experience and observation, that a researcher will opt for re-eliciting over searching for the data. In this way, valuable time with the consultant can be misused on a repeated task and the opportunity to ask the next question is postponed and, sometimes, missed altogether.

Exacerbating the problem of needless re-elicitation is the challenge of inter-researcher data-sharing. For the same reasons as discussed above, researchers involved in theoretical, documentary or pedagogical endeavours would significantly advance the progress of their projects if they had easy access to each others' primary linguistic data.

Often when data are shared between researchers, that data is of insufficient quality for the purposes of the receiving researcher. Without notes on contexts of use, the semanticist will find little use for the lists of transcriptions and glosses of the syntactician; and without high quality audio recordings and the ability to efficiently extract minimal pairs, the phonetician will find little use for the data of either.

Linguistic researchers with widely varying goals, theoretical assumptions and methodologies can find significant value in one another's data. The increase in efficient use of such data is particularly crucial when the languages under study are endangered and as fluent speakers become ever more rare. In order to meet the three main challenges described above and increase the efficient use of primary linguistic data, I have developed the BLD as a Blackfoot-specific prototype of a more general and object language-independent piece of (open source) software called the Online Linguistic Database (OLD). In the next sections I discuss the features of the BLD which help to improve efficiency and the current state of the OLD project.

¹ Funding for this project provided by a SSHRC Image, Text, Sound and Technology Grant awarded to M. Wiltschko. Funding for its presentation at the 2009 Algonquian Conference provided by a SSHRC Standard Research Grant awarded to E. Ritter.
The BLD: Facilitating Efficient Use of Blackfoot Language Data.

The Blackfoot Language Database is a web application that enables Blackfoot researchers to store their text, image, sound and video data in a secure, highly searchable and collaborative online system. The BLD provides structure and powerful search functionality, thereby aiding researchers in finding appropriate data quickly. It is also a multi-user web application with password protection and as such it permits immediate sharing of linguistic data between authorized researchers. Finally, the BLD model facilitates the storing and sharing of data in a variety of formats – textual transcriptions, audio and video recordings, and textual and audio-visual representations of stimuli/contexts – and in doing so provides a rich source of data that can be maximally re-used for purposes different from those originally intended.

![A screenshot of the BLD. At present (Oct 24, 2010), the BLD contains 4,750 Forms, 2,109 Files and 109 Collections.](image)

The basic unit of data in the BLD is a Form. A Form is a morpheme, word, phrase or sentence. Forms consist of an obligatory transcription and gloss and the following optional fields: morpheme break, morpheme gloss, general comments, speaker comments, method of elicitation, morpho-syntactic category, keywords, grammaticality and further metadata such as date of elicitation, speaker, elicitor, etc. This structure gives users the option of restricting their search patterns to specific fields, thus improving searchability.

The BLD enables the following methods of search: substring match, exact match, substring match on any or all of a space-delimited set of strings, regular expression match and a set of forced choice restrictor expressions. The regular expression functionality, in particular, allows for fine-grained searching. Regular expressions are a language for expressing complex patterns using sets, boolean operators, quantification and the ability to refer to the beginning and end of the string. Using regular expressions and the search functionality of the BLD, one could, for example, search for all transitive verbs (i.e., those with vta or vti as their syntactic category) that are bisyllabic and start with i.

Being an online application, the BLD can be accessed by a registered researcher wherever an internet connection is available. As soon as researchers add their textual and multi-media data to the database, it is immediately available to all other researchers. Requiring authorization to access the BLD mitigates against the possibility that such inter-user data-sharing might betray the confidence of those speakers (and their communities) who do not want their data to be made wholly public.
In addition to the Forms described above, the BLD contains two other basic entities: Files and Collections. Files are textual documents, digital images, audio or video along with textual metadata. Collections are stories, discourses or transcriptions of elicitations and their primary content consists of ordered lists of Forms. Files can be associated to both Forms and Collections. Thus a Form may be associated to a File that is an audio/video recording of a speaker's utterance and also to a File that is a pictorial/video stimulus that was employed in the elicitation of the Form. Similarly, a Collection may be associated to a File that is an audio recording of an entire elicitation or an entire story. As the BLD allows in-page embedding of Files and their media, a user may, as an example, read a Blackfoot story with interlinear analysis and simultaneously listen to audio recordings of each sentence, pausing and replaying as desired.

The ability to upload media files as BLD Files and associate them to Forms and Collections makes the system an easily accessible storage place for high quality linguistic data. Far beyond the barebones data of transcriptions and glosses, the BLD allows the storage of a variety of textual information, as well as multimedia data consisting of recordings and stimuli. In this way the BLD enables the repurposing of linguistic data.

At present, the BLD contains over 4,700 Forms and 2,100 Files grouped into 109 Collections all contributed by 30 researchers from the universities of British Columbia, Calgary, Lethbridge and Montana. The database continues to grow and as it does its value as a resource for linguistic researchers will increase. Researchers interested in accessing the BLD are encouraged to contact the author.

THE OLD: SOFTWARE FOR CREATING LINGUISTIC WEB APPLICATIONS,
I have argued that primary linguistic data, i.e., the transcriptions, recordings and stimuli generated by linguistic researchers qua fieldworkers, are used inefficiently and that a collaborative linguistic web application can help to increase efficiency by improving searchability, enabling easy inter-researcher data-sharing and providing a structure for organizing multimedia linguistic data that increases their capacity for repurposing.

The Online Linguistic Database (OLD) is a piece of software that allows groups of researchers to create a web application (akin to the BLD) for collaboratively documenting their language of study. The OLD is open source software licensed under the GPL (http://www.gnu.org/licenses/gpl.html).
The OLD provides all of the functionality implemented in the BLD, plus the following notable additions: (i) a dictionary-like interface to the data, (ii) automatic and configurable conversion between orthographies and (iii) functionality for exporting to a variety of formats, including plain text, tab-separated values and XeTeX.

The home page of the OLD is http://www.onlinelinguisticdatabase.org. This page provides links to the OLD user guide, the Python Package Index page from which the OLD can be downloaded, the Google Project Hosting page where the OLD source code can be found and the OLD Google Groups discussion page.

The OLD home page is also the front page of a demo OLD web application. To interact with the Demo OLD, go to the OLD home page and follow the instructions to login.

Persons interested in using, playing with, contributing to or discussing the OLD are encouraged to (i) contact the author and (ii) explore the Demo OLD.

**SUMMARY.**

I have argued that efficient use of primary linguistic data can have significant effects on the progress of language documentation, pedagogy, revitalization and analysis. I have presented the Blackfoot Language Database (BLD) as a tool that can increase efficiency by improving searchability, promoting data-sharing and facilitating the repurposing of linguistic data. The Online Linguistic Database is open source software that allows researchers to easily create a collaborative documentation tool like the BLD. It is my hope that tools such as these will be beneficial to linguistic fieldworkers and those whose work depends on access to primary linguistic data.