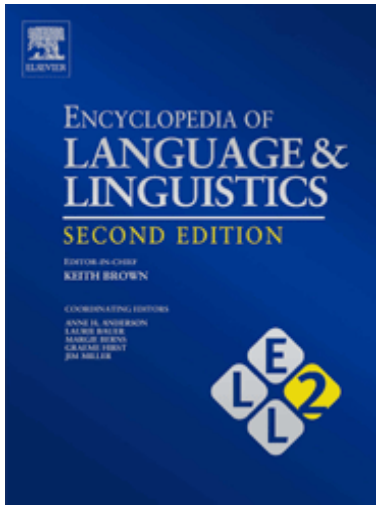

Thieberger N (2006), Computers in Field Linguistics. In: Keith Brown, (Editor-in-Chief) *Encyclopedia of Language & Linguistics, Second Edition*, volume 2, pp. 780-783. Oxford: Elsevier.

Implications for technologies to support remote collaborative work.' In Hinds P & Kiesler S (eds.) *Distributed work*. Cambridge, MA: MIT Press. 137–162.
Lee J (2004). 'A BlackBerry throbs, and a wonk has a date.' *New York Times Sunday Styles, Section 9*, May 30. 1–2.
Ochsman R B & Chapanis A (1974). 'The effects of 10 communication modes on the behavior of teams during cooperative problem-solving.' *International Journal of Man–Machine Studies 6*, 579–619.
Ohaeri J O (1998). 'Group processes and the collaborative remembering of stories.' Unpublished doctoral dissertation, State University of New York at Stony Brook.
Schober M F & Clark H H (1989). 'Understanding by addressees and overhearers.' *Cognitive Psychology 21*, 211–232.
Whittaker S (1995). 'Rethinking video as a technology for interpersonal communications: theory and design implications.' *International Journal of Man-Machine Studies 42*, 501–529.
Whittaker S (2002). 'Theories and methods in mediated communication.' In Graesser A, Gernsbacher M & Goldman S (eds.) *The Handbook of Discourse Processes*. Hillsdale, NJ: Erlbaum. 243–286.
Whittaker S J, Brennan S E & Clark H H (1991). 'Coordinating activity: An analysis of interaction in computer-supported cooperative work.' In *Proceedings of CHI '91: Human Factors in Computing Systems*. New Orleans, LA: Addison-Wesley. 361–367.
Williams E (1977). Experimental comparisons of face-to-face and mediated communication. *Psychological Bulletin 16*, 963–976.

Computers in Field Linguistics

N Thieberger, The University of Melbourne, Melbourne, Victoria, Australia

Computers have been associated with field linguistics from their earliest days, as witness the enthusiasm with which computers were embraced by linguists, from mainframe computers in the 1960s to personal computers in the 1980s. While initially it was common to force our efforts into the framework provided by particular software, we are now more aware of the need to see the data itself as the primary concern of the analyst and not the software that we use to manipulate the data. Inasmuch as it allows us to carry out the main functions desired by a field linguist, software is a tool through which our data passes, the data becoming transformed in some way, but surviving the journey sufficiently to live on, independent of any software, into the future.

In this article, I discuss ways in which computers can assist field linguists whose chief concerns I take to be language documentation, including recording a previously unrecorded or little recorded language in order to write a grammatical description.

Field linguistics has been going through a change in focus over the past few years. There is increasing recognition of the need to record languages with few speakers, and to support such speakers with materials such as text collections, dictionaries, and multimedia (e.g., text, audio, images, and video). Computers are central to this effort, especially as we move to digital recording in which there will be no analog original. Laptop and palm computers are common tools for the first-world linguist, as are solid-state digital recorders and digital video cameras, which produce digital files for access on computers. Processing power of computers keeps increasing as does storage and RAM, which means we are now able to deal with real-time media (audio and video) in ever larger quantities, raising crucial issues for data management.

A typical workflow engaged in by a field linguist is presented below, together with a description of methods for working with small and perhaps endangered languages, and for managing the data so that it can be analyzed. Further analytical tools, like morphological parsers, are considered in the article on Natural Language Processing (NLP) (*see* **Natural Language Processing: Overview**).

An interest in supporting endangered languages, and the efforts of speakers or their descendants to learn about them, encourages us to focus on archival methods and on producing the best quality material for access in the future. Thus, the focus here will be on computer-based tools for analyzing linguistic material in ways that allow it to be safely stored, retrieved, and reused by others, as discussed by Bird and Simons (2003) in a work that is central to the present discussion.

For the linguistic fieldworker, the usual workflow involves recording, transcribing, and interlinearizing a corpus so that there is a base of information for analysis. This analysis is written as a grammar and may be accompanied by a collection of texts and a dictionary of the language. There may also be a set of media files that are linked to by their transcripts, allowing readers to hear audio or see video in the

language. In addition, this material is housed in a suitable repository, a digital archive which preserves the data for future use.

The types of tasks that we will need to carry out in the analysis of a previously unrecorded language are outlined below. Assuming that we begin with recordings (digital, or analog converted to digital) that are the primary data, we first need to label them clearly, so that they are identifiable from the moment of recording, and to establish a database of metadata, the who/what/where/when information that is easily forgotten in a short time without good descriptive notes. It is useful at this stage to have considered a naming convention, so that the tapes can be permanently identified in both our own documentation and in any archive in which we lodge the data. (Filenames should persist over time so that any reference to them can be resolved, for example by someone looking through the data in the future. Filenames should not contain unusual characters that various computer systems find difficult to recognize.) Maintaining a good database of the items (tapes, transcripts, texts, images, etc.) and of the relationships between them allows us to keep track of derived forms and the context from which they are derived.

We then need to transcribe the media to produce a textual index in whatever form we require. Transcription can be undertaken with tools that capture time-alignment, so that the resulting file has timecodes associated with chunks of text.

We should be clear from the outset that we are engaging in a data management task, in which complex relationships between types of ethnographic data need to be tracked, both for our own use of them and for assisting in retrieving information in the future. Database structures can assist here, but only if they do not lock up the data in a proprietary format (one that is owned by a company rather than being 'open source' or publicly and freely available). Relational databases allow us to reflect relationships in the data and to avoid duplication by listing, for example, items on a tape linked to the names of speakers and their characteristics (age, sex, etc.), and the derived information (such as texts, media files, and lexicons).

In the late 1980s, Lancashire (1991) listed a number of software tools for various aspects of linguistic analysis, many of them aimed at working with large corpora of metropolitan languages. Not all of these are useful from the point of view of a fieldworker recording a small language (one with relatively few speakers and typically with no written record), as the programs deal with what we can characterize as 'high-end' applications such as NLP or analysis based on very large datasets.

An issue that was dealt with extensively in the late 1980s was representation of orthographic typefaces by fonts, and it may not be too optimistic to say that we are about to overcome these problems by means of the international standard, Unicode, in which most character sets have found a home. While field linguistics is not addressed as a subject heading in Lancashire's compilation, more recent work by Johnston (1995) and Antworth and Valentine (1998) is devoted to just this topic and surveys the relevant software of the time in some detail.

Some of the tools described in these two sources are still used by field linguists, but this is partly because there is no choice. Shoebox is an example of a fine piece of software that is the mainstay of lexicographic and textual analysis and was last updated in 2000, although it has recently been replaced as Toolbox on Windows platforms. A number of tools have not been updated and are now unable to run on recent operating systems.

Bearing in mind that the data is our primary concern and not the software we use to manipulate it, it is nevertheless critical that the software enables us to perform the kinds of tasks we routinely require in order to assist us in our fieldwork. It is the function of a software tool to transform data, or to allow us to interact with the data. We take it as given that the tools discussed here may soon be superseded. The kinds of functions that we need as linguists will continue to be addressed in new ways in the future. As there is no one tool that will do all that we require, we need ways of allowing our data to flow between the tools. This typically involves the use of text manipulation software or regular expression parsers.

Most of the examples of tools listed below can be found on the Internet, and searching for the major headings here will locate any more recent items. There is an enormous possibility for new uses of linguistic data, both in the exploration of its internal links and in the representation of the data itself, to accompany our analyses or to assist in language reintroduction programs. Given that this is the case, it would be foolhardy to suggest that we could provide all the answers in a fixed time or location. Rather, there are major sources of information on these topics, as given in the list of web links below, that should be consulted by anyone wanting to locate current information on these topics. They should also get in touch with the local linguistic archive that will be keeping abreast of the best emerging practices.

Transcribing

Producing a textual index (or transcript) of a media file, with timecodes inserted into the resulting file.

Elan, http://www.mpi.nl/tools/elan.html
Transcriber, http://www.ldc.upenn.edu/mirror/Transcriber/
Clan, http://childes.psy.cmu.edu/clan/
TASX, http://tasxforce.lili.uni-bielefeld.de/

(cf. the Annotations page which has a list of many of these kinds of tools: http://www.ldc.upenn.edu/annotation/)

Interlinearizing Text

Providing an annotation of the transcript, in a morpheme-level correspondence, typically with reference to a controlled vocabulary that will become a lexicon of the language.

Shoebox, http://www.sil.org/computing/shoebox/
Toolbox, http://www.sil.org/computing/toolbox/

Building a Corpus of Media Material

Amassing transcripts linked to media files to allow navigation through the media via the textual index. Instantiating links established with transcription tools.

Audiamus, http://www.linguistics.unimelb.edu.au/thieberger/audiamus.htm

Concordancing the Corpus

Establishing a list of all words in the corpus in their context. Ideally this concordance interacts with the corpus to allow you to move between the concordance and the corpus (McEnery and Wilson 2001: 209ff., give a list of tools for corpus research).

Conc, http://www.sil.org/computing/conc/
Wordsmith, http://www.lexically.net/

Conversion of Linguistic Data

To restructure our data for use in the tools listed here we need conversion methods that can take the data from one format to another. Regular expressions allow the linguist to query the data on structure rather than content. So, for example, the expression '\r.' will find any carriage return and following character, regardless of what it is. Similarly, '\r[0-9]' finds any numeral in that position. Regular expressions assist in structuring textual data to move it between applications. A general search on 'regular expression' will give more information, see for example http://www.regular-expressions.info. Tools that use regular expressions include:

Emacs, http://www.emacs.org
BBEdit, http://www.barebones.com/products/bbedit/index.shtml
Perl, http://www.perl.com
ECONV, (http://www.mpi.nl/tools/econv.htm) does conversions between *Shoebox*, *Transcriber* and *Elan* textual formats without the need to learn regular expressions.

Building a Dictionary Based on the Corpus

Shoebox, http://www.sil.org/computing/shoebox/
Databases programs are, in general, not recommended for building dictionaries as they are too restrictive on the form in which an entry can be represented. A major benefit of Shoebox is that it provides a means for glossing texts linked to a dictionary, a function that is not available with other tools. Dictionary presentation tools are a useful way of getting structured lexical information into a public form, for example:

Kirrkirr, http://www-nlp.stanford.edu/kirrkirr
LexiquePro, http://www.lexiquepro.com

Spectral Analysis

Acoustic analysis of segments of field recordings can be accomplished with these two widely used tools.

Praat, http://www.fon.hum.uva.nl/praat/
Emu, http://emu.sourceforge.net/

Archiving Data

These archives are both repositories for field recordings and derived forms of data and analysis and clearinghouses for relevant information on linguistic methods and tools.

Digital Endangered Languages and Musics Archive Network (DELAMAN), http://delaman.org/
Open Language Archives Community. (OLAC), http://www.language-archives.org/
Aboriginal Studies Electronic Data Archive (ASEDA), http://www.aiatsis.gov.au/rsrch/rsrch_pp/ased_abt.htm
Archive of the Indigenous Languages of Latin America, http://www.ailla.utexas.org
Documentation of Endangered Languages (DOBES), http://www.mpi.nl/DOBES
Endangered Languages Archive (ELAR), http://www.hrelp.org/archive/
Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC), http://paradisec.org.au

Linguistic Computing Directories

General sources of information on linguistics and computing tools.

http://www.sil.org/linguistics/computing.html
http://www.linguistlist.org/sp/Software.html

See also: Character Sets; Natural Language Processing: Overview; Phonetics: Field Methods; Semantics: Field Work Methods.

Bibliography

Antworth E & Valentine R J (1998). 'Software for doing field linguistics.' In Lawler J & Dry H A (eds.) *Using computers in linguistics: a practical guide*. London; New York: Routledge. 170–196.

Bird S & Simons G (2003). 'Seven dimensions of portability for language documentation and description.' *Language 79*, 557–582.

Johnston E C (1995). 'Computer software to assist linguistic field work.' *Cahiers des sciences humaines 31(7)*, 103–129.

Lancashire I (1991). *The humanities computing yearbook 1989–90*. Oxford: Clarendon Press.

Lawler J & Dry H A (eds.) (1998). *Using computers in linguistics: a practical guide*. London; New York: Routledge.

Leech G N, Myers G & Thomas J (eds.) (1995). *Spoken English on computer: transcription, mark-up, and application*. Harlow, Essex, England; New York: Longman.

McEnery T & Wilson A (2001). *Corpus linguistics: an introduction*. Edinburgh: Edinburgh University Press.

Computers in Lexicography

A Kilgarriff, Lexicography MasterClass, Brighton, UK

Computers can be used in lexicography to support the analysis of the language and to support the synthesis of the dictionary text. There are, of course, many other interactions between computing and lexicography, including the preparation and presentation of electronic dictionaries, the use of dictionaries in language technology systems (*see* **Computational Lexicons and Dictionaries**), and the automatic acquisition of lexical information (*see* **Controlled Languages; Lexical Acquisition**). They will not be covered here.

In technologically advanced dictionary-making, the lexicographer works with two main systems on their computer: the corpus query system (CQS) for analysis and the dictionary writing system (DWS) for synthesis. Currently, these are always independent, with communication between the two via cut and paste. We describe requirements, and the state of the art, for each.

Dictionary Writing Systems (DWSs)

Anyone producing a dictionary needs to (a) write it, and (b) store it. Each can be done on either paper or computer. 'Dictionary writing system' means the software used where either or both are done on a computer.

Producing a dictionary is a large and complex operation. The DWS can facilitate the operation at many points. Dictionary production usually involves a team whose members include lexicographers, a chief editor, a project manager, and a publisher. The DWS will be a key tool for all of them, each from a different perspective. The lexicographer wants the tool to facilitate writing and editing text. The chief editor wants it to support quality checking and consistency, including ensuring that dictionary policies are observed. The project manager wants it to support progress monitoring, including the process of allocating packages of work to lexicographers, distributing them, and checking that they are returned on time. The publisher wants it to deliver a versatile database that can readily be used for producing various dictionaries (electronic and paper, large and small) and potentially for licensing for a range of other purposes, such as spell-checking or automatic translation.

The Dictionary Grammar

A dictionary is a highly structured document. An entry typically contains a headword, pronunciation and part-of-speech code, optional labels, and information about inflectional class and morphological and spelling variants, then a sequence of senses, each with definition or translation and optional examples. Each of these is a different information field.

There are constraints on which fields are required or allowed where. Fields are often distinguished by font or use of bold or italics. Some fields, like part of speech, may only take one of a small set of values; others play a specific role in sorting or cross-referencing. A lexicographer or user of an electronic