develops a generalizable approach to integrating communities without internet access in the ongoing curation of digital language materials they have produced.

We are particularly interested in the possibilities afforded by the increasingly common presence of mobile communication technologies in areas where basic infrastructure such as electricity, running water, and even roads remain absent. In such areas, new social institutions involving cell phones have begun to evolve, for example, conventions for exchanging cell-phone minutes, or new gathering spaces defined by signal access. Exploiting this new form of technology, the non-profit organization Open Mind has developed a system called 'Question Box' (see `http://questionbox.org/about-mission`) which allows remote communities to access information in critical domains such as health, agriculture, and business. Google's voice-based social media platform SayNow is being used to allow cell phone users in Egypt and elsewhere to leave voicemail messages that appear online immediately as Twitter audio feeds (`http://www.nytimes.com/2011/02/02/world/middleeast/02twitter.html`). We believe that similar methods are worth exploring as a means to creatively connect less-networked language communities with researchers and archives.

References

Christen, K. (2008). Archival Challenges and Digital Solutions in Aboriginal Australia. *SAA Archeological Recorder* 8(2): 21-24.

Czaykowska-Higgins, E. (2009). Research Models, Community Engagement, and Linguistic Fieldwork: Reflections on Working within Canadian Indigenous Communities. *Language Documentation and Conservation* 3(1): 15-50.

Dobrin, L. M. (2008). From Linguistic Elicitation to Eliciting the Linguist. *Language* 84: 300-324.

Yamada, R.-M. (2007). Collaborative linguistic fieldwork: Practical application of the empowerment model. *Language documentation and conservati on* 1(2): 257-282.

Language Documentation and Digital Humanities: The (DoBeS) Language Archive

Drude, Sebastian
Sebastian.Drude@mpi.nl
Max-Planck-Institute for Psycholinguistics, The Netherlands

Trilsbeek, Paul
Paul.Trilsbeek@mpi.nl
Max-Planck-Institute for Psycholinguistics, The Netherlands

Broeder, Daan
Daan.Broeder@mpi.nl
Max-Planck-Institute for Psycholinguistics, The Netherlands

1. Overview

Since the early nineties, the on-going dramatic loss of the world's linguistic diversity has gained attention, first by the linguists and increasingly also by the general public. As a response, the new field of language documentation emerged from around 2000 on, starting with the funding initiative 'Dokumentation Bedrohter Sprachen' (DoBeS, funded by the Volkswagen foundation, Germany), soon to be followed by others such as the 'Endangered Languages Documentation Programme' (ELDP, at SOAS, London), or, in the USA, 'Electronic Meta-structure for Endangered Languages Documentation' (EMELD, led by the LinguistList) and 'Documenting Endangered Languages' (DEL, by the NSF). From its very beginning, the new field focused on digital technologies not only for recording in audio and video, but also for annotation, lexical databases, corpus building and archiving, among others. This development not just coincides but is intrinsically interconnected with the increasing focus on digital data, technology and methods in all sciences, in particular in the humanities.

As one result, a number of worldwide and regional specialized language archives have been established, devoted to a new type of corpora: digital, multimedia, multi-purpose. The DoBeS archive alone contains data on more than 60 languages; it is hosted at the Max-Planck-Institute for Psycholinguistics (MPI-PL) in Nijmegen, where it is combined with data from other field research projects, in total covering around two hundred languages. The Technical Group of MPI-PL (now as a new unit, 'The Language Archive', TLA) has not only been developing new tools such as ELAN and the language archiving technology suite, but is also active in building regional archives offering the same technology around the globe, and in networking with other archives, for instance in DELAMAN.

Furthermore, The Language Archive (TLA) also participates, based on our experience with the DoBeS archive, in international initiatives such as CLARIN, which have a much broader scope than endangered languages. Still, the new type of corpora, including the DOBES archive, have only incipiently been explored with novel research questions and methods, and the full potentials of cooperation between language documentation and, e.g., computational linguistics and automatic data processing methods have only begun to be exploited. This also holds for topics such as community member participation and representation, engaging wider audiences, access methods, ethical issues around privacy and 'publicness', documentation and archives' relations to social networking and other mass participation platforms. In general it seems obvious that neighbouring disciplines can benefit from language documentations, and that also language documentation and archiving can gain much from cooperation with similar activities and research in other fields, which may even change the way language corpora are designed and future documentation projects are carried out. Nevertheless, the concrete details have still to be determined in the quickly growing environment of 'digital humanities'.

2. Outline of the DoBeS programme and its relation to the digital humanities

The DoBeS programme emerged from long discussions between the Volkswagen foundation and a small group of linguists around the recently founded German 'society for endangered languages', concerned with language diversity and linguistic field research. First result of this interaction was a successful summer school about 'language description and field research' in Cologne in 1993, a very early response to the first public statements urging to get active about the problem of 'endangered languages', i.e. the imminent loss of humanities linguistic (and cultural) diversity. To that time the very first uses for the internet started to be developed, such as the world wide web; email was rarely used in business and academia, university term papers were still often written on a typewriter or by hand. Although the beginning of 'computing and the humanities', 'linguistic computing' etc. actually dates back long before the arrival of personal computers (at least to the 1970ies, when todays major DH associations were founded), the digital humanities (DH) as we know them today existed in the early nineties maximally in an embryonic state, for neither sufficient digital data sets nor the needed computational tools and methods existed.

The first call for applications for individual projects of what would soon become to be called the DoBeS programme was made in 1999. Despite the long period of intense debate and planning, it was still all but clear how language documentation (LD) exactly would be carried out. Himmelmann had already published his seminal paper 'Documentary and descriptive linguistics' (1998), but still the exact goals of such a research program and especially its methodologies had to be decided, tested and established, and the needed technology had to be identified, designed and developed. Thus, the first year of DoBeS was designated to be a pilot phase, where a consortium of linguists, leading eight different documentation projects on a variety of languages all over the globe,[1] together with the technicians of the technical project at the MPI-PL, would discuss and eventually decide on fundamental, methodological, technical and legal/ethical issues, giving clear guidelines to the next generations of projects and establishing the fundaments of a technical infrastructure to support the building of digital language corpora of a new kind – focussed on actual oral language use in a natural cultural context.

By the beginning of the programme in 2000, the general technological setting had changed dramatically. Semi-professional audio and video equipment with satisfactory quality for serious documentation work were now available to affordable prices (although to that time, video recordings were often still analogue, e.g. in HI8 format, and audio recordings were often done in a compressed format, e.g. ATRAC). Most importantly, digital storage capacities had grown to a point where even video recordings with a reasonable resolution and bit-rate could be stored on usual hard discs and even on removable media such as DVDs. This allowed LD to be a fundamentally digital enterprise, aiming at the building of large digital language archives, i.e., multimedia corpora with natural speech data.

At the same time, the technical group (TG) at the MPI-PL had started to tackle the problem of the increasing amount of digital data produced and used at that institute. Some of the data obviously were of relevance for the future – for re-use with different research goals, or just for being able to check and reproduce the results, making the research more accountable. The TG had already started to work on a digital archive of research data (including data from linguistic field research) and was thus in the best position to function as the technical centre for the DOBES programme, which in turn for several years boosted the development at MPI-PL. Most importantly, a meta-data-schema, IMDI, was developed with decisive input from the DOBES consortium of the pilot phase and in the first years of the main phase (most of the 8 projects of the one-year pilot phase also participated in the first years of the main phase of the DOBES programme). Also, the development of a multi-media annotation tool, ELAN, started. Other tools and infrastructure elements were added over the years to what by around 2008 came to be called the 'Language Archiving Technology' suite.

These developments occurred basically without connection to the first developments in the mainstream 'digital humanities' or 'E-Humanities'. In the 1990ies, when larger data collections became available, when individual computers became part of every university department, and when the development of tools tailored to specific needs was easily done, research techniques and tools such as text mining, quantitative text analysis, complex databases were increasingly often employed by technophile humanities practitioners and computer linguists. Language documentation, however, was developed by a completely different community of field researchers (linguists, anthropologists, music-ethnologists etc.) which usually were not akin to digital technologies and computational methods. The object of study had little overlap, too: computer linguistics and the emerging digital humanities in general were (and continue to be) mainly concerned with major languages and predominantly in their written form, whereas LD by definition is concerned mainly with small and understudied languages, most of which are only occasionally written.

So for about almost a decade, the two areas developed mostly independently one from another. The DoBeS program grew (each year between five and ten new projects started, each with a duration of usually three to four years) and had followers (see above) and became more mature, as the basic standard methodologies were clear and new research questions were introduced and a stronger interdisciplinary approach consolidated. The necessary tools became available, more stable and increasingly easy to use. In particular, Language Archiving Technology with the web-based programs LAMUS for the upload and archive integration, AMS for archive management, and other IMDI or ELAN related tools were built, so that the DoBeS archive (at the core of the larger digital archive with language resources at MPI-PL, 'TLA') became an example for digital archives.

In the last years, the two communities and research traditions (DH & LD) have begun to come closer. In linguistics, language documentation has contributed to raise the interest in linguistic diversity, linguistic typology and language description. This general movement also affects computational linguistics which now increasingly shows also interest in small languages. At the same time, language documentation has from its beginning been concerned with digital data and methods, even if mostly for data management and archiving and less for linguistic analysis. Still, teams for LD projects nowadays usually have quite good computational expertise, and field workers are less distant from digital tools and data than they used to be, and than many other linguists. Furthermore, soon it became obvious that the time was ripe for novel research questions and topics that would make use of these linguistic corpora of small languages, which is where computational linguistics and other statistical methods come into play (see below).

In the opposite direction, the digital humanities grow and consolidated to the point of engaging in constructing major research infrastructures for their needs, integrating the numerous individual data sets and tools 'out there' in the many departments and individual computers, and allowing humanities research to be carried out on a completely new higher level. The DH projects such as DARIAH and in particular CLARIN count with the experiences at the MPI-PL, in particular with the DoBeS archive. The Language Archive is now one of the backbone centres in CLARIN and participates actively in developing this DH infrastructure in Germany, the Netherlands and in a European and international level.

With a total of about 70 major project funded, the DoBeS programme is in its final phase now (the last call for projects was in 2011), and already it has been one of the most successful programs of the Volkswagen Foundation, in terms of impact and public awareness. Internationally, LD has grown into a respected sub-discipline of linguistics and neighbouring fields on its own right, as is witnessed by a successful on-line peer-reviewed journal (Language Documentation and Conservation, LD&C) and a bi-annual international conference (ICLDC, 2009 and 2011 in Hawaii) that is attended by many hundreds of participants from all over the world. In this community, the DoBeS programme is generally recognized as trendsetter and in several aspects as a model, and there are numerous personal and technical links between DoBeS and other LD initiatives (ELDP, DEL, EMELD, PARADISEC, etc.).

3. Some key issues of Language Documentation and Digital Humanities

One distinctive feature of The Language Archive is diversity, on different levels. First, it is concerned with the very linguistic diversity inner- and cross-language, uniting data from many different languages with unique features and a broad variety of communicative settings. But the data, produced by many different teams with different background and research interests, are also diverse in their formats and contents – what is annotated, and how it is annotated. While with respect to metadata and archiving, DoBeS was successful to create agreement and consensus among the researchers, the same does not hold with respect to levels and conventions of annotation, from the labelling of 'tiers' in ELAN or other annotation tools to the abbreviations used for grammatical glosses and labels. This now constitutes a major obstacle for advanced cross-corpora research, and even with the general ISOcat data category registry, much manual work has to be done before different corpora are interoperable. The same holds for lexical data, as the work in the RELISH project has shown which created interfaces and conversions between different standards for lexical databases (LMF & Lexus and LIFT & LEGO).

Another issue is the question of sustainability. Still, too often one finds great initiatives that produce wonderful tools and/or data sets, but without any long term plan – when the funding ends and/or the developer leaves, the resources are abandoned and not rarely unusable after some years, when hardware and software changes. There are different aspects to sustainability – one is the sheer preservation of the bit stream, which is threatened by eroding media such as hard disks or optical discs. The necessary automatic copying of several backups to different locations and the constant replacement of out-phased hardware can only be done by data centres, with which smaller archives should cooperate. The Language Archive was lucky enough to be able to negotiate a 50 year guarantee for bit stream preservation by the Max Planck Gesellschaft already around 2006. For data format accessibility, one needs to rely on a manageable number of open standards (such as XML and UNICODE for text data, or widely used open and preferably not compressed codecs for audio and video data), and has to be prepared to migrate the whole corpus from one format to another if new standards supersede the ones used in the archives (although the original files should always be preserved, too). Finally, the problem of maintenance of tools is generally not satisfactorily solved. Due to changing hardware, platforms, drivers, standards, most tools are bound to need constant maintenance even if no new features are to be implemented (which usually happens if a tool is well received by a large user community), and few funds are available for this kind of activity. One has to be constantly considering how and with how many resources the currently offered software can and should be maintained or further developed. All these questions are by now well known in the DH and now addressed by the infrastructure projects such as DARIAH and CLARIN, but have been addressed at a comparatively early stage at The Language Archive.

It is interesting to notice that one of the strengths of The DoBeS Language Archive is its connecting character. Not only are the data in the archive relevant for many different disciplines, not just linguistics, but also anthropology, history, psychology, music-ethnology, speech technology etc. Also some of the tools developed at The Language Archive are now more widely used, in particular ELAN, which is now used not just by descriptive and documentary linguists, but fostered multi-modality (gesture) and sign language research and is even employed in completely unrelated fields. Also ISOcat has the potential to bridge the gap between different traditions of labelling entities in areas much broader than linguistics, and ARBIL, the successor of the IMDI metadata-editor, is by now being prepared to be the major tool for the creation of modular and flexible CMDI metadata as used in CLARIN. Finally, other archives using Language Archiving Technology have been and continue to be set up at different locations in the whole world, constituting a network of regional archives which soon can interact, exchange their data for backup and ease of access purposes, and hence strengthen and consolidate not just LD but generally the archiving of valuable digital research data.

There are several big challenges ahead for language documentation and the corpora it produces. Some, as the tension between the general movement to open access to research data and the need to protect the individual and intellectual property rights of speakers and researchers are in principle solved by providing different levels of (controlled and possibly restricted) access and employing codes of conducts and other ethical and legal agreements. Still, this state of affairs has constantly to be re-thought and discussed. The same holds for new insights and methods for language strengthening and revitalization, promoting multilinguality and the fruitful interaction between coexisting cultures, and the digital inclusion of linguistic, social and cultural minorities. Obviously, these are questions that cannot be solved by science (alone).

Others are only beginning to be properly addressed, such as the mobilization language data: the future shape of language documentation and archives will be radically different from its current state, in view of new opportunities, needs and goals beyond data gathering, language description and revitalization. How can be ensured that many users, from researchers via the speakers themselves and the general public, make the best use of the data, that novel research questions are addressed and answered with the support from language documentation corpora? How can the process of creating, annotating and archiving new, high-quality documentation data be made easier even for community members without the presence and support of a LD project and researchers from abroad?

The closer interaction of LD and computational techniques as being developed in the context of the DH will certainly help to improve the situation with respect to one of the major impediments in LD: the high costs for annotating the rich corpora. This is currently done mostly by hand, albeit in some cases with semi-automatic support, such as in the case of morphological glossing based on string-matching with a lexical database. Still, segmenting recordings into utterances, identifying the speaker, transcribing the original utterance and providing further annotation is a very time-intensive work which is mostly done by experts and only to a smaller part can be delegated for instance to well-trained native speakers. Here new computational methods which work reasonably well for written major languages can be generalized or adapted for other languages, and statistical methods can be used to segment and label audio and video data based on speaker and/or gesture recognition. The team at The Language Archive is working on the inclusion of such methods into their tools. The better this integration is, the broader are the possible uses in many humanities disciplines way beyond the field of language documentation.

Notes

[1] Three projects on geographically close languages in the Upper Xingu area in central Brazil formed a collaborative project within the consortium.

The potential of using crowd-sourced data to re-explore the demography of Victorian Britain

Duke-Williams, Oliver William
o.duke-williams@ucl.ac.uk
University of Leeds, UK

This paper describes a new project which aims to explore the potential of a set of crowd-sourced data based on the returns from decennial censuses in nineteenth century Britain. Using data created by an existing volunteer based effort, it is hoped to extract and make available sets of historical demographic data.

Crowd-sourcing is a term generally used to refer to the generation and collation of data by a group of people, sometimes paid, sometimes interested volunteers (Howe 2007). Whilst not dependent on Web technologies, the ease of communication and ability to both gather and re-distribute digital data in standard formats mean that the Web is a very significant enabling technology for distributed data generation tasks. In some cases – the most obvious being Wikipedia – crowd-sourcing has involved the direct production of original material, whilst in other cases, crowd-sourcing has been applied to the transcription (or proof-reading) of existing non-digitised material. An early example of this – predating the Web – was Project Gutenberg (Hart 1992), which was established in 1971 and continues to digitize and make available texts for which copyright has expired. A more recent example in the domain of Digital Humanities is the Transcribe Bentham project (Terras 2010), which harnesses international volunteer efforts to digitize the manuscripts of Jeremy Bentham, including many previously unpublished papers; in contrast to Project Gutenberg, the sources are hand-written rather than printed, and thus might require considerable human interpretative effort as part of the transcription process. Furthermore, the Transcribe Bentham project aims to produce TEI-encoded outputs rather than generic ASCII text, potentially imposing greater barriers to entry for novice transcribers.

FreeCEN[1] is a project which aims to deliver a crowd-sourced set of records from the decennial British censuses of 1841 to 1891. The data are being assembled through a distributed transcription project, based on previously assembled volumes of enumerator's returns, which exist in physical form and on microfiche. FreeCEN is part of