Workshop "Potentials of Language Documentation: Methods, Analyses, and Utilization"

Leipzig, 3-4 November 2011

Lecture Hall, 2nd floor

Thursday, November 3		
08.30 - 09.00	**Registration**	
09.00 - 13.00	**PANEL 1: Methods** Chairs: Frank Seifart, Peter Wittenburg, Daan Broeder	
09.00 - 09.15	Introduction to Panel 1	
09.15 - 09.35	E-Grammars and endangered languages corpora	Sebastian Drude
09.35 - 09.55	Corpus linguistics' perspective on language documentation data and the challenge of small corpora	Anke Lüdeling
09.55 - 10.15	Language assignment in DoBeS and similar corpora of endangered languages	Jost Gippert
10.15 - 10.35	Using R as an interface to DoBeS corpora	Balthasar Bickel
10.35 - 10.55	Interim summary and discussion	
10.55 - 11.15	Coffee	
11.15 - 11.35	Prospects for language comparison	Michael Cysouw
11.35 - 11.55	Unsupervised Morphological Analysis of Small Corpora	Amit Kirschenbaum, Peter Wittenburg, Gerhard Heyer
11.55 - 12.15	Supporting language research with generic automatic audio analysis	Daniel Schneider, Oliver Schreer
12.15 - 13.00	Summary and discussion Panel 1	
13.00 - 14.30	Lunch	
14.30 - 18.30h	**PANEL 2: Analyses** Chairs: Anna Margetts, Geoffrey Haig, Nikolaus Himmelmann	
14.30 - 14.45	Introduction to Panel 2	
14.45 - 15.05	Information structure, variation and the referential hierarchy	Jane Simpson
15.05 - 15.25	Data from language documentations in research in referential hierarchies	Stefan Schnell
15.25 - 15.45	On the social determinants of linguistic complexity	Peter Trudgill
15.45 - 16.15	Interim summary and discussion	
16.15 - 16.45	Coffee	

16.45 - 17.05	The interpretation of frequency in corpora and its consequences	Sabine Stoll
17.05 - 17.25	Bilingualism and multimodality	Marianne Gullberg
17.25 - 17.45	Tours of the Past through the Present	Marian Klamer
17.45 - 18.30	Summary and discussion Panel 2	
19.30	Workshop dinner (Auerbachs Keller)	
Friday, November 4		
09.00 - 13.00h	**PANEL 3: Utilization** Chairs: Dagmar Jung, Paul Trilsbeek	
09.00 - 09.15	Introduction to Panel 3	
09.15 - 09.35	Online presentation and accessibility of endangered languages data: The General Portal to the DoBeS-archive	Gabriele Schwiertz
09.35 - 09.55	Online presentation and accessibility of endangered languages data	Hans-Jörg Bibiko
09.55 - 10.15	Uses of language documentation, data format and accessiblity, and social media	Nick Thieberger
10.15 - 10.45	Interim summary and discussion	
10.45 - 11.15	Coffee	
11.15 - 11.35	Language Archives: They're not just for linguists anymore	Gary Holton
11.35 - 11.55	From language documentation to language planning: not necessarily a direct route	Julia Sallabank
11.55 - 12.15	Creating educational materials from language documentation data	Ulrike Mosel
12.15 - 12.30	Evolving uses in the context of social media	Paul Trilsbeek & Dagmar Jung
12.30 - 13.00	Summary and discussion Panel 3	
13.00 - 14.30	Lunch	
14.30 - 18.00	**Concluding discussion**	
	Summarizing statements by panel coordinators	
	Summarizing statement by Bernard Comrie	
	Summarizing statement by Ulrike Mosel	

Abstracts

Sebastian Drude

E-Grammars and endangered languages corpora

There are several ways in which modern corpora of language use can be connected to digital grammars, mostly depending on what is understood by 'digital grammar' (or 'e-grammar'). I will present and briefly discuss three current directions of research.

If one understands under digital grammar a traditional scholarly text in digital form describing in prose the structure of a language, possibly enhanced by hypertext features (hypertext grammar), then a language corpus is not only part of the empirical basis for the analysis, but also the source for examples, which can be linked to the relevant sections of the hypertext grammar. Search mechanisms can be used to provide more examples then those selected by the author, if the language data is sufficiently annotated. The tags and labels used in glossing can in turn refer to the respective sections of a grammar. Christian Lehman, Sebastian Nordhoff and Sebastian Drude, among others, are pursuing these possibilities.

A second approach is to extract typological information from a set of glossed texts while building a database of syntactically and partly morphologically annotated sentences. This new field of research combines several new trends such as grammatical engineering (grammar Matrix project), treebanks and parsers of glossed digital texts as produced by language documentation projects. Although applying a general model, this approach promises to provide highly significant and accurate analytical typological information on languages, even with only a small corpus available. It also enhances the data corpus by adding syntactic trees to the sentences. This approach is currently promoted by Emely Bender and others.

A third current line of research is developed by Mike Maxwell and others. It combines aspects of the previous two, advocating a new style of grammar writing which is akin to what is known as "literate programming" (Donald Knuth) in software developing, particularly in the open source community. The idea is to combine the descriptive prose text with more formalized "code" which represents in a standardized and abstract way the (phonological and morphological) structure of the language. This code can be translated in instructions for an automatic parser which then can check the underlying data set (all example sentences, or an entire transcribed corpus, even without morpheme glosses) for accuracy and completeness of the rules.

All three approaches are still in an experimental stage, but the first explorative results are impressive and point at future possibilities which twentyfive or even fifteen years ago would not have been imagined by grammaticographers.

Anke Lüdeling

Corpus linguistics' perspective on language documentation data and the challenge of small corpora

While each language/variety has, of course, its own problems that have to be addressed specifically for that variety there are general issues that have to be solved for all languages/varieties. [1]

In my statement I will focus on two such issues, illustrating my statements with examples from the German learner corpus Falko (Lüdeling et al. 2008). The first issue pertains to the primary data; the second deals with annotation.

[1] The common distinction between languages with many resources (English, German, French etc.) and languages with few resources (less-known languages, under-resourced languages etc.) is not helpful because it implies that the problems are solved for the first group of languages and they somehow intrinsically differ from all the other languages.

- Many parameters (such as medium, purpose, socio-economic factors, etc.) influence the way we speak or write, and ideally these parameters should be relevant in corpus design. Some of these parameters have been well-known for years and reflect in corpus design, typically on a fairly abstract level (spoken/written, academic/personal etc., cf. Hunston 2008). But specialized corpora, deeply annotated data and better statistical models[2] show that there are many more and more fine-grained categories that play a role in influencing variation than previously thought. This means that we need (a) small specialized corpora of as many varieties as possible, (b) a way to analyze different corpora at the same time and (c) the possibility to add metadata at any time.
- Even for languages with an established orthography and a long tradition of standardization there is an enormous degree of variation and tools and resources are typically only able to deal with a very small portion of the varieties (mostly formal, written varieties). This often means that 'non-standard' varieties are not annotated at all.[3] I argue that it can be helpful to use a standard annotation scheme even for non-standard varieties because standardization is a matter of degree.

References
Biber, Douglas (2006). University Language. John Benjamins, Amsterdam.
Hunston, Susan (2008) Collection strategies and design decisions. In: Lüdeling, Anke & Kytö, Merja (eds) Corpus Linguistics. An International Handbook. Mouton de Gruyter, Berlin, 154-168.
Lüdeling, Anke; Doolittle, Seanna; Hirschmann, Hagen; Schmidt, Karin & Walter, Maik (2008) Das Lernerkorpus Falko. In: Deutsch als Fremdsprache 2/2008, 67-73.
Reznicek, Marc & Golcher, Felix (2010) Stylometry and the interplay of topic and L1 in the different annotation layers in the FALKO corpus. Paper presented at QITL-4 (Quantitative Investigations in Theoretical Linguistics), Berlin.

Jost Gippert

Language assignment in DoBeS and similar corpora of endangered languages

The paper deals with the question of language assignment in DoBeS and similar corpora of endangered languages.

Given that recorded texts of endangered languages abound in code-switching (mostly between the endangered vernacular and dominant languages, but also other languages involved in the bi- and multilingual settings that are typical for endangered languages), the distinction of the different layers present in the texts may be crucial for all kinds of (language-specific or cross-linguistic) research in these corpora, as well as for the theory of language endangerment in general. For the time being, the annotation schemes provided in the DoBeS, framework do not admit of an easy differentiation of linguistic units pertaining to different linguistic layers, and language-specific search functions are still wanting.

The paper discusses ways to cope with this, considering, among other things, the advantages of the emerging standard of ISO 639-6 ("language names"; cf. www.geolang.com/iso639-6/).

[2] E.g. reproducible register differences between certain parts of academic texts (Biber 2006) or topic-effects within the same register (Reznicek & Golcher 2011).
[3] Sometimes tools are specifically written (or re-trained) for a 'non-standard' variety. This makes it difficult to compare results across varieties and, in essence, only creates another 'standard'.

Balthasar Bickel

Using R as an interface to DoBeS corpora

Fieldworkers are used to look at corpora in the format of interlinearly glossed text. In my presentation I will argue that it is useful to transform corpora into tabular format where each column corresponds to an annotation tier and each row to a clause (or whatever is one's favorite unit of analysis). A tabular format allows convenient manipulation of the data (e.g. extracting stems from complex forms, combining subsamples etc.) and is fully compliant with modern tools for statistical analysis and data visualization, such as R (R Development Core Team 2011)

I will illustrate the advantages of this approach by way of a case study on the acquisition of ergativity in Chintang (Sino-Tibetan, Nepal; Stoll et al. 2011). A popular hypothesis in language acquisition is that children learn new categories in an item-based fashion (Tomasello 2003): the combinatorial potential of the new category with lexical items is small in the beginning and reaches adult-like patterns only gradually. The Combinatorial potential of a category can be modeled as its lexical information entropy. Entropies in turn can be directly visualized and explored statistically. I will show that, once the corpus is transformed into tabular format, the relevant computations are extremely easy, using only a handful of commands in R.

References

R Development Core Team, 2011. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing,http://www.r-project.org.

Stoll, S., B. Bickel, E. Lieven, G. Banjade, T. N. Bhatta, M. Gaenszle, N. P. Paudyal, M. Rai, N. K. Rai, & I. P. Rai, 2009. Audiovisual corpus on the acquisition of Chintang by 6 children (ca. 350,000 words). DOBES Archive, http://www.mpi.nl/DOBES.

Tomasello, M., 2003. Constructing a language: a usage-based theory of language acquisition. Cambridge, Mass.: Harvard University Press.

Michael Cysouw

Prospects for language comparison

The comparative study of language has mostly been based on relatively limited information in the form of grammatical descriptions and (small) lists of words. The prospect of using larger amounts of electronic corpora opens up many new fascinating prospects and possibilities for language comparison.

There are two aspects of corpus preparation that I would like to stress in the view of comparative analysis, namely (a) multilateral parallelism and (b) incremental annotation. Both aspects are related to the usability for comparative analyses, but also to enhance the development of automatic processes to assist the descriptive linguist.

With multilateral parallelism I would like to stress the utility of including a limited set of relatively 'artificial' translated utterances for purely comparative purposes. The basic proposal here is to include some (short) texts to be translated in the course of language documentation, e.g. the "North Wind and the Sun" story, maybe the Universal Declaration of Human Rights, or some specific texts as prepared by linguists for comparative purposes. Such data is purely intended as a "normalization point" for comparative purposes. However, it might also be beneficial for the development of tools for automatic analysis as the same text will become available in many languages in parallel.

As for corpus preparation, there are of course numerous possible wishes as for the structure and the annotation. However, I think is it important to realize that language comparison can already benefit from rather low-level prepared corpora. The central insight here is that, for a language-specialist, every minute detail is of importance for a better understanding of the language in

question. In contrast, for the comparative linguists it is already possible to compare languages on the basis of limited understanding as well. I would therefore like to propose the following levels of corpus-preparation and annotation as steps of increasing detail. Most importantly, it should be realized that there are many intermediate steps between raw data and completely glossed data that are already usable. So, it might be considered to prepare parts of a corpus only to a limited extend to reduce the work load. Specifically, I see the following incremental steps as intermediates to full glossing:

1) utterance-by-utterance translation (note that this step might be performed before transcription!)
2) transcription (already a practical orthography is very useful; full phonetic detail is not necessary always)
3) stemming (intended here is a process of approximately identifying lexical roots/stems only, without necessary glossing them; think of this step as forming set of wordforms in the corpus that are based on the same root/stem)
4) morpheme separation (i.e. the identification of morpheme boundaries without necessarily identifying there precise meaning of the morphemes identified)
5) morphemic annotation ("full glossing")

Amit Kirschenbaum, Peter Wittenburg, Gerhard Heyer

Unsupervised Morphological Analysis of Small Corpora

Present unsupervised methods for linguistic analysis of language data are typically developed for large corpora and usually perform rather poorly on small corpora of a few hundreds of thousands or even just thousands of words. In this presentation, we introduce an outline of an interactive system of unsupervised and supervised algorithms that is design to assist the annotation of both large and small corpora. It shall be language independent and produce a high quality annotation of languages with various properties (e.g. with both concatenative and non-concatenative morphology, various word orders etc.) The system shall be developed within a frame of a joint project between the University of Leipzig and Max Planck Institute in Nijmegen.

The unsupervised module of the system attempts to compensates the data sparseness problem by using the linguistic information of various sources (morphological, POS, semantic levels) and by taking advantage of the interaction between these levels. As a result, the module produces suggestions for linguistic analyses on the three levels, which the annotator can manually correct. The supervised module is then trained on the corrected data and produces the final annotation of the corpus. This system of interactive annotation shall be integrated into the existing and widely used environments such as ELAN, LEXUS, or VICOS to make the annotation process as efficient as possible.

The focus of the presentation is on the morphology analysis method that shall become a part of the unsupervised system. It first calculates significant co-occurrences between words, within a context of a sentence. Co-occurrence statistics are further used as features to determine distributional similarity between words. Next, for each word orthographical similarities between the word and the set of its distributionally similar words is obtained, from which the list of the closest ones, based on normalized edit distance, is selected. The underlying rationale for creating such a list is that such words may be derivations or inflections of the same stem.

The input word and its corresponding list are then aligned by a multiple sequence alignment that is typically used in bioinformatics to lay out a pair of sequences of e.g., proteins or genes.

The algorithm is adjusted and enhanced to effectively deal with linguistic data and detect both morphological patterns in a given language and morpheme boundaries of its tokens. Preliminary results for both large corpora (German) and small corpora (German and Kilivila) will be introduced.

Daniel Schneider, Oliver Schreer

Supporting language research with generic automatic audio analysis

In this presentation, I will describe two generic scenarios where automatic audio analysis can support language research: speed-up of annotation and acoustic query-by-example.

1. Automatic analysis can speed-up the annotation process and unleash human resources, which can then be spent on theorizing instead of tedious annotation tasks. I will describe selected automatic tools that support the most time-consuming steps in annotation, such as segmentation, speaker detection and time alignment of existing transcripts.

2. With automatic acoustic query-by-example, researchers can search for acoustic phenomena of interest in a large non-annotated corpus, and retrieve similar occurrences that might support their theory. Query-by-example can be used as a tool to unlock large, non-annotated corpora, where complete and perfect automatic annotation is not feasible yet.

Jane Simpson

Information structure, variation and the referential hierarchy

Silverstein (1976)'s hierarchy of features and ergativity (referential hierarchy) was proposed to capture apparent systematic variation in the linguistic/morphosyntactic expression of the grammatical functions Subject and Object and the linked semantic roles Agent and Patient. The variation was expressed in terms of binary features (e.g. +/- 1st singular) and markedness (absence of morphological marking correlates with naturalness of association - e.g. animate agents acting as subjects are less likely to have overt case marking than inanimate agents). This was connected with a hierarchy with pronouns at the top and nouns at the bottom, which suggests that the basis of the referential hierarchy lies in information structure.

Better language documentation allows cross-linguistic testing of the descriptive adequacy of the referential hierarchy. The data Silverstein used came mostly from case-suffixing Pama-Nyungan languages (Dyirbal) and verb agreement languages (Chinook). Counter-examples have been noted, including (by Silverstein) the Pama-Nyungan language, Arrernte, where counter to expectation first person singular shows a three way distinction (Ergative, Nominative and Accusative), whereas the other pronouns show Nominative-Accusative pattern.

Better language documentation also allows better understanding of the basis of the referential hierarchy in a given language. For example, an assumption of the original hierarchy was obligatoriness of marking, rather than optionality. Optionality (i.e. choice of marker or its absence) is often associated with a different or additional semantic/pragmatic force. This in turn may determine reanalysis and subsequent change in the linguistic/morphosyntactic expression of Subject/Object/Agent/Patient. Along the way, apparent counter-examples to the referential hierarchy may be created. I discuss this with respect to Arrernte and Central Australian Pama-Nyungan languages.

Stefan Schnell

Data from language documentations in research in referential hierarchies

I will highlight two aspects of language documentations that proved highly relevant in the investigation of referential hierarchies in Vera'a. (1.) the culture-specifics and discourse-structural workings of animacy and referentiality only really emerge in contexts of sufficient richness and comprehensiveness of an accessible text corpus of indigenous genres; elicitation has obvious

drawbacks as it fails to elude the particularities of culture-specific animacy-concepts (and often imposes SAE concepts) and lacks the discourse context inevitable for thorough treatment of referentiality patterns. (2.) relevant referentiality patterns are often manifested in statistically relevant patterns across discourse, rather than absolute categorical rules, something that simple "hierarchies" cannot really capture.

I will illustrate these two points with two concise examples from my work on the Vera'a language, namely the classification of nominal expressions in terms of animacy on the basis of their morphosyntactic behavior in different types of construction, and the preferred realization of P-like arguments.

Peter Trudgill

On the social determinants of linguistic complexity

Linguistic complexity developed in societies of intimates. It was in such societies that the major complexity-producing social factors were maximally operative: small size, dense social networks, large amounts of shared information, high stability, and low contact (Trudgill *Sociolinguistic typology: social determinants of linguistic complexity.* OUP, 2011). Dixon says "the most complex grammatical systems … are typically found in languages spoken by small tribal groups". It is possible that with the disappearance of societies of intimates, we will also see the disappearance of complexifying linguistic changes, especially as the social matrices afforded by these societies provide the sociolinguistic conditions which *permit* complexity development; but they don't compel it. There may also be a trend towards a predominance of simplifying changes – Wray & Grace say "a language that is customarily learned and used by adult non-native speakers will come under pressure to become more learnable by the adult mind, as contrasted with the child mind" – and thus in the long run a significant reduction in overall world-wide linguistic complexity.

Bickerton argued for the importance of the development of creoles by children in high-contact situations as a window onto linguistic competence. But - in a kind of mirror-image argument – I suggest that if we want to learn more about the inherent nature of linguistic systems and their propensities, we must actually focus attention on low-contact varieties. Isolated languages may have, to those of us of a European-language background who speak standard creoloids and koinés, amazing and unusual features; but these are important since they represent the limits to which languages can go when, as Bailey says, they are "left alone". Wohlgemuth suggests that "rarities" are more likely to occur in languages with small numbers of speakers, which are therefore more likely to be endangered – another reason for arguing that more linguistic fieldwork should urgently be carried out before such features are, perhaps, lost to linguistic science for ever.

A while back I was talking with an eminent generativist. I asked how he would handle switch reference in his current model. He replied: "I don't know. That's something you only get in exotic languages. I don't know anything about exotic languages." One implication was that if a phenomenon occurs only in a small far-away language which is exotic to an academic speaker of a standard European language, it's not worth bothering about. In fact, these "exotic languages" are not exotic at all. They are normal. This is what human languages must have been like throughout most of the tens of thousands of years of human history.

If we want to learn more about the nature of linguistic systems, we should direct our attention to the structures and changes that occur in the dwindling number of low-contact, dense social network languages in the world. We must hurry, not only because most of the world's languages are in danger, but because most of those that are going to be left behind will tend to be of a single, historically atypical type. We can assume that the human language faculty has remained unchanged for very many millennia. But the sociolinguistic matrices in which linguistic changes occur have changed significantly; and we are increasingly unlikely ever again to see the development of highly inflectional, fusional language varieties; or languages with 80 consonants, or 31 personal pronouns, or seven-term evidential systems. Linguists of the future may well find themselves envying linguists of today the opportunities that we still have for the study of highly complex languages at first hand.

Sabine Stoll

The interpretation of frequency in corpora and its consequences

One of the main advantages of the recent development of corpora of endangered and small languages is the possibility to do quantitative research also in those languages for which otherwise only a grammar and a lexicon and may be a few texts would have been collected. These corpora now allow us to make quantitative statements about the frequency of constructions and this changes our possibilities of the types of linguistic analyses we can do in these languages drastically. However, quantitative analyses need to take into account a number of different issues before they can be performed. I illustrate these issues by an analysis of ergative case distributions in Chintang (Sino-Tibetan, Nepal), using a large acquisition corpus of four children and their surrounding adults. I will show that frequency is too vague a concept to be of use and that we need instead very precise measurements, tailored to the specific research question at hand and the available sampling space.

Marianne Gullberg

Bilingualism and multimodality

It is often said that most people in the world speak two or more languages and that bilingualism (or multilingualism) should therefore be considered the norm rather than the exception. As such, (adult) second language acquisition (SLA) and bilingualism should be at the heart of research agendas focused on the nature of linguistic systems and language use in context. Yet they are not. Moreover, although all studies of SLA and bilingualism are inherently concerned with crosslinguistic and typological variation (and make the sweeping claim above), they generally draw only on data from Western European (standardised) languages studied in classroom settings. Virtually nothing is known about acquisition in other contexts. There are therefore enormous gaps in our knowledge about adult acquisition and bilingualism in other settings that have consequences for the validity of the theorising around these issues. Conversely, studies of language contact, shift, and loss rarely consider theoretical linguistic claims in SLA and bilingualism concerning acquisition, language processing, and linguistic systems in general.

In a similar vein, we know that language acquisition and use is situated in a context of multimodal behaviour, making the study of gesture, for instance, fundamentally important. Yet, the multimodal nature of language acquisition and language use is sorely under-explored. Studies of gestures predominantly focus on practices in Western European languages. Only a few studies have looked at how speech, gestures, practical actions, and artefacts are mobilised and orchestrated in a wider set of languages. And again, although language documentation now often comprises multimodal resources such as audio and video, these are rarely exploited to address theoretical questions regarding the multimodal nature of language, acquisition, and bilingualism.

Clearly, vital theoretical and empirical gains could be made if researchers in these fields collaborated and considered data and frameworks from language documentation. I will suggest some possible avenues of investigation to highlight the importance of language documentation on theorizing, and conversely, ways in which current analytical work may influence documentation.

Marian Klamer

Tours of the Past through the Present

Language documentation efforts result in pictures of the Present that always feed into pictures of the Past. Which Past exactly, depends on (i) the type of synchronic **data** we analyse, and (ii) the **area** we study. Also, the resulting histories differ in (iii) the type of contact **scenarios** proposed, and (iv) the supposed **time** depth.

I present illustrations from Eastern Indonesia, my field of expertise. Linguistically and ethnically, Eastern Indonesia constitutes the interface between the Austronesian and Papuan worlds. Papuans have lived in this area for more than 40,000 years, whereas Austronesians came down from Taiwan less than 6,000 years ago. The area was originally Papuan but became largely "austronesianised" through the incoming Austronesians who assimilated with the original populations, though in some locations, Papuan languages are still spoken. In his 1995 Introduction to the *Comparative Austronesian Dictionary*, Darrell Tryon noted that Eastern Indonesia was 'perhaps the least known area in the Austronesian world'. Since then, we have gone from knowing almost nothing to having a body of documentation on approx. 10% of the 200 languages in this area. These data fed into historical and contact linguistics in many different ways, of which the following are just three illustrations.

1. After the initial stage of focusing on descriptions of single languages, a natural next step was to synthesize across a larger sample of languages, in order to get a more detailed historical profile of the area. Research questions included: How can we say which language is Austronesian and Papuan in Eastern Indonesia? Which features are Austronesian, which are Papuan, and if languages have shared or mixed features, are they the result of shared inheritance or contact? The **data** on which the analysis was built were scattered typological data taken from Austronesian and Papuan languages in a huge geographical **area** (over 1000 km West-East); and the proposed contact **scenarios** remained vague, including notions such as diffusion through subtrate contact between Austronesian and Papuan, taking place in **prehistoric** times.

2. As more details became available about particular regions within Eastern Indonesia, more specific puzzles could be addressed. One such puzzle was the existence of Austronesian speakers on the coasts of Alor and Pantar who are considered 'newcomers' by their Papuan neighbors, while no-one can explain where they came from and when. To investigate the history of this group, the language was compared with its closest relative in the **area** some 200 kilometers away, as well as with 5 Papuan neigbor languages. The kinds of **data** used included survey word lists and morphological and syntactic data that had become available only a couple of years before, as well as historical and ethnographic notes. The contact **scenarios** involved immigration, trade and marriage exchange, with an estimated **historical** time depth of 500-600 years.

3. Presently ongoing research studies language contact at the level of individual bi-lingual speakers, in whose brains two languages co-exist. In Eastern Indonesia, every speaker of a local minority language (L1) also speaks Indonesian (L2). In studying language contact at the individual we consider specific **data** such as the use of numeral classifiers. Can we attest language change in bilinguals whose L1 lacks numeral classifiers, but whose L2 has obligatory classifiers? In this kind of contact study the **area** is a single village; the contact **scenario** involves L1-L2 connections in the brains of individuals, and the time depth is **two generations**.

Gabriele Schwiertz

Online presentation and accessibility of endangered languages data: The General Portal to the DoBeS-archive

When considering the exploitation of language documentation data contained in language archives, three major user groups can be identified: The speaker community, the scientific community, i. e. linguists and scholars of related disciplines and the general public. Each of these user groups has different interests and different needs all of which are hardly satisfied by the IMDI-tree representation of the archive. For the community users, community portals have been created in some projects. However, in order to increase traffic in the archive, facilitate access to the data and generate new user communities and scenarios, a general portal for non-community members has been created for the DoBeS-archive.

This portal tries to address the needs of the different user groups with different sub-portals where selected information specific to typical queries of those users is presented in a straightforward way and shortcuts to the relevant media are included. In the future, this portal will serve as the main entryway to the archive and as a container of information about the DoBeS-projects.

This talk will present features of the portal and the underlying intentions. It will put to discussion potential usage scenarios and how to best accommodate future user requests. It is hoped that such a portal will help to broaden the utilization of the collected data in the archive especially by non-DoBeS-members.

Hans-Jörg Bibiko

Online presentation and accessibility of endangered languages data

In my talk I would like to show some possibilities how language data can be visualized. By using maps in order to display words in different languages of a semantic concept offers the user the chance to see and learn the geographical distribution of used words for the same meaning in a specific area. Furthermore it is relatively simple to generate an online dictionary out of Toolbox files. One possible software environment to create such maps or online dictionaries is "R" [www.r-project.org] since the EAF file format (XML) and Toolbox files (many thanks to Taras Zakharko for writing this R-plugin) can be read and processed by "R" quite easily. Finally I would like to show an interactive application to show e.g. relations between different languages. This kind of interaction can attract people esp. the younger generation much easier because they can "play with language data" and hopefully they'll begin to ask questions.

Nick Thieberger

How can language documentation data be utilized in a broader context?

There are obviously many ways in which language data can be used (by speakers and their descendants as heritage mementos, or for language learning; by other researchers), but first it must be created in ways that permit ready reuse. That is, the data must exist in formats that allow it either to be used immediately or to be converted to a useable form without too much effort. This implies that we have tools that produce output in reusable forms and that linguists have training in what it means to create reusable data. Currently only recipients of DoBeS and ELDP funds are required to deposit in these archives, while others are welcome to use the same facilities, but, in general, do not. All other language documentation activities need to be brought into this same regime if possible, using the network of existing archives. Making it easier to access the archive and enter metadata will increase uptake, but there need to be citation methods and rewards for those who establish good

collections arising from their fieldwork. We have developed an online system for presenting interlinear text and media (EOPAS.org) that can be used to display short edited stories.

How must these digital data be stored, represented, and made accessible by the archives?
Accessibility is based on locatability of the material in the collection. This relies on good catalogs using standard terms and accessible via normal search mechanisms (e.g., google, Open Archives Initiative). Non-compliant repositories, those whose catalogs do not conform to the normal standards, should be encouraged to conform. For repositories that will not be able to conform (that is, state archives or similar institutions), a service could be built that indexes language material in these collections.

Once an item has been located it should be available for use, if possible. Examples are the DOBES data sets, or our online collections of papers by Capell, Wurm and Roesler. By early 2012 PARADISEC will provide streaming access to most of its collection.
Accessibility also implies that analog material is digitised. While newly created linguistic records are typically digital, a great deal of legacy material exists only in analog forms and so is outside of the scope of much current language archive infrastructure.

What kinds of uses will evolve in the context of the social media?
The uses of data can be: (1) online and (2) local or offline. For online use of data there must be persistent location and identification that allow citation and resolution of links. This means we need proper repositories with longterm commitment to curating the material. Socal media can play a role in dissemination or publicity, but the critical factor is the longevity of the data itself. Once people start combining data from disparate sources they will create new research objects that themselves need to be identified and curated. A new research environment needs to be created to deal with what could be 'mashups' or could involve correlating transcripts and media, or images of handwritten transcripts and media.

Offline use is likely to be most relevant to speakers of endangered languages. Data formats are crucial here too, as it must not be too difficult to convert from the archival form of an item to a deliverable form. For example, a dictionary of a language should be derived from a structured lexical data set, as in Toolbox. Similarly a set of texts for production in a book can be derived from a set of interlinear glossed texts in Toolbox. Media for a CD or DVD can also be readily converted to playable formats and used in ITunes installations.

Gary Holton

Language Archives: They're not just for linguists anymore

While many language archives were originally conceived for the purpose of preserving linguistic data, these data have the potential to inform knowledge beyond the narrow field of linguistics. Today language archives are being used by people without formal linguistic training for purposes not necessarily envisioned by the original documenters. In this paper I describe two such non-linguistic uses which are becoming increasingly important at the Alaska Native Language Archive.

The first such use involves the archive as a source of cultural documentation. Language documenters are first and foremost field workers, interacting with speaker consultants whose interest lies in the documentation of many types of knowledge, be they marriage traditions or navigational techniques. A field worker documenting names for kin terms or stars is very likely to also document knowledge of marriage customs or stellar navigation, respectively. Furthermore, much language documentation has been collected without regard to the nature of non-linguistic content. A text may be recorded because it represents a particular genre or style, such as narrative or conversation. The content of that recording—i.e., what that narrative or conversation is about—is generally not constrained by the documenter. As a result language archives now present a veritable treasure trove of non-linguistic information encoded in the signal of the subject language.

A second important use involves the creation of secondary language materials based on archival data. By far the most active user community at the Alaska Native Language Archive consists of descents of speakers of Alaska Native languages. Heritage language communities seek a diverse range of materials—recordings, word lists, place names maps, stories, songs—and often assemble them in unique ways. Language communities may also interested in associated information such a photographs and materials on bilingual education. To the extent that language archives hope to support revitalization of the languages which their holdings document, the archives must be prepared to assemble materials in new ways.

This presentation will provide examples of both of these "non-linguistic" uses of language archives. Recognizing the way these materials are being used will help us to better plan for utilization of language documentation resources.

Julia Sallabank

From language documentation to language planning: not necessarily a direct route

In this presentation I will consider how documentary linguists can provide support for community language planning initiatives, and discuss some issues. These relate partly to the process of language documentation: what and who we choose to document, and how we define 'a language'; and partly to community attitudes and dynamics.

Language planning falls into two main categories:

- Actions to define or modify a language itself – especially *corpus planning*
- Actions to modify *attitudes* towards a language, or its *status* in a language ecology.

The most obvious area in which documentary materials and linguists have a role to play is corpus planning: codification, graphisation, orthography, standardisation, terminology development. Documentary materials provide evidence-based corpora for the production of dictionaries, grammars and language learning materials.

Language maintenance and revitalisation are often grass-roots initiatives, and it is at this level that documentary linguists are best placed to provide support But in community-led language planning, documentary evidence may not be appreciated (or believed, or taken on board) by 'purists' or 'traditionalists'. There may be unwillingness to accept the fact of language change or variation, and possible antipathy to those who point it out (such as linguists). In addition, orthography in particular can generate heated debates related to deep-seated language ideologies.

Language planning is notorious for being ineffective, so archived material may be seen as a fall-back position in case there is desire for revival – as has happened in Europe, America and Australia even 200 years after the 'last speakers'. This necessitates "a record of a language which leaves nothing to be desired by later generations wanting to explore whatever aspect of the language they are interested in" (Himmelmann 2006: 7; Woodbury 2003). But what is 'representative'? What and whose language do we represent, and who decides? I argue that community-based language planning is situated in community dynamics. This raises the question of how we define 'language community', who 'owns' a language, and how and which members decide on community language policy?

Examples of such dilemmas from our documentation will be given. Language documentation and language planning may be taking place simultaneously, in which case I argue that part of our task is also to document the process of language planning itself.

For community-based planning, archives need to be locally available, bearing in mind that by no means all users of the materials will be internet users – for reasons of lack of resources or age. So who do we aim these materials at? There are similar constraints on social media: they motivate younger users and learners, but people who use such media are unlikely to be traditional users of endangered languages. If a computer with projector and internet access are available I will show examples.

Ulrike Mosel

Creating educational materials from language documentation data

This paper briefly describes the workflow of the production of a school book series in the Teop language, their content and their educational purposes. The documentation project started with audio recordings of spontaneously told narratives, but when the Teop team members saw the trasncriptions, they wanted to edit them before publication and produce books. By doing the transcriptions and the editorial work, the Teop team members have turned into creative writers and presently work on descriptions of the natural enviroment.

The project is a grass-roots project. We did not involve any official authorities of the provincial government, but only worked together with teachers of the local village school. In the meantime two books have officially been launched, and the village school has been awarded a prize for their initiative. Apart from a hymn book printed in the fifties of the last century, these books are the only books in the Teop language.

Since the Teop people do not have regular access to electricity, TV, DVD players etc. we did not produce electronic materials for the schools. But all data, including PDFs of all texts, the sketch grammar and the lexical database, are stored in the DoBeS archive. For the next years we plan to conduct workshops for teachers, but we do not know yet, where to get funds.

For linguistic research the Teop Language Corpus represents a unique resource for the investigation of differences between spoken and written (edited) language, and the development of new genres and registers in the course of a language documentation project.