Language documentation in a globalised world

Oddrun Grønvik

Oddrun Grønvik
Hovudredaktør, Norsk Ordbok 2014.
e-post: oddrun.gronvik@iln.uio.no

1 Why document languages with a view to creating standard languages?

Why is it an important task to document vernaculars with a view to creating written standard languages? There are a number of answers to this compound question, ranging from 'understanding an interesting system' through 'preserving a unique knowledge and interpretation of the world' to 'necessary for education and development'. All of those answers are true enough. They are also similar, in that they emphasise the outsider angle - the view the rich part of the world often has of the less (economically) developed.

The individual right to language can be assumed to spring out of the UN Declaration of Human Rights, which states that "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers." (UDHR Article 19). It is obvious that this freedom "to seek, receive and impart information and ideas" presupposes mastery of a written standard language which is used in all walks of life in the linguistic community the individual happens to be part of. Information offered in a (written) language that is not understood or mastered as a means of expression, can just as easily become a tool of repression.

Large sections of the declaration of individual human rights can only be realised for the individual if the communities that the individuals are part of, also have their rights respected. One of those rights has to be the right of a community to express itself through its own language and define and present its own culture through the sets of concepts underlying the language and the culture. This is essential both in order to develop a sense of collective identity, and in order to facilitate economic empowerment.

Why does the mere existence of a written standard language bring a sense of empowerment to a linguistic community? This is best understood if one looks at what a standard language is. A modern standard language represents a measure of agreement and acceptance within a linguistic community on all the essential aspects of language description. The essential aspects include choice of alphabet; analysis and description of phonology, morphology, word creation and syntax; and finally core vocabulary, as an index to the register of concepts that shape thinking and communication within the language community itself. The existence of a known and used written standard language is a strong statement from the linguistic community that it exists as a community, and that it has sufficient self-awareness to describe and define itself.

To a language community the state of possessing a standard language is an entirely different state from being a community with a vernacular that is studied by outsiders, but not defined by the community itself. A language community which has and feels ownership to a written standard for its mother tongue, also has much better means of creating and mastering the self-image that the community presents to the world.[1] A language community with a vernacular and a spoken culture may be equally sophisticated, but it lacks the possibility to store and present its culture in writing, and runs a real risk of being seen by the outside world only as an object of more or less exotic study. Defining a culture and a community through language also means drawing the borders against other linguistic communities - not in an inimical fashion, but in order to say: "This is who we are, and this is how we present ourselves through our language".

Of the roughly 100 national languages that are in general use in their countries, six languages are what the Summer Institute of Linguistics (SIL) terms "international languages".[2] Of the remaining languages out of the 7000 total, many have a good deal of literature and documentation, but even if the existing documentation has a high academic standard, it will quite often be piecemeal and accessible only to other linguists who are interested in the same group of languages. Ph.D. dissertations do not necessarily lead to the production of school grammars.

A majority of the languages of the world will go largely undocumented for a long time unless a more efficient way is found of documenting

1 Cf.f.i. the UZ Proposal To Establish an African Languages Research Institute1999 :1 "These activities are of national significance, particularly since they are raising people's consciousness about language issues, their own language situations and the context of language policy formulation."

2 Chinese, Spanish, English, Hindi, Arabic and portuguese are also termed world languages (Prys Jones 1998 p. 302 f).

them, as a first step towards establishing them as standard languages. Given the expense and the amount of work required in both initial and further language description, it is rational to look for ways of improving efficiency without sacrificing quality. Let us therefore look at the issue of task, process and method and in light of conclusions reached on these issues, the question of who should be the preferred work force – the mother tongue linguist or the field linguist from abroad.

2 Factual documentation

Linguistic standards must build on documented facts about the language, and those facts should be accessible to all users. The need for language documentation corresponds to the need for documentation on any subject requiring public management and policy. In order to know what is there, one needs to take stock, organise and document. This is accepted as obviously true for f.i. natural resources, health, agriculture and finance. The same should go for language.

Documenting a language that only exists as a vernacular is a huge task, no matter who takes it on. Success depends on good and knowledgeable informants, skilled field linguists and a fruitful cooperation between the documenting parties, in addition to time and money. It is a task requiring patience and trust, as well as skills and resources. Once the initial documentation is established and has yielded a (preliminary and tentative) orthography, there remains the task of collecting a body of text which is large enough to provide a foundation for analysing and describing syntax, and extracting and defining the vocabulary.

Creating a written standard language is expensive. In European history there are many examples of outstanding endeavour from individuals in the creation of written standards. But a closer examination shows that written standard languages are not created through lone work in libraries and chambers of study, nor are they established by outsiders to the linguistic community in question. All the established European standard languages were first documented as standard languages by mother tongue linguists. Editing Dr. Johnson's Dictionary of the English language (1755) took eight years, required a specially outfitted building and six helpers in addition to the chief editor[3], and was funded by a group of London booksellers throughout. The project basis – English literature after the reformation in the 16th century and up to the time of production – represents a huge body of work which had to be evaluated and excerpted. The Norwegian linguist and lexicographer Ivar Aasen spent more than thirty years documenting the Norwegian vernacular, the last ten on his major dictionary of the proposed standard language (Aasen 1873. Preface p. XVIII).

3 British Library: http://www.bl.uk/learning/langlit/dic/johnson/1755johnsonsdictionary.html

Aasen wrote every word of his manuscripts himself, but he did not work in a vacuum. In the preface to the dictionary he expresses his gratitude to all supporters and helpers[4], whom we know to have been many, and a great deal of those years was spent travelling round the country on field excursions. He also made use of older glossaries of the Norwegian vernacular, to the extent their contents could be verified. Most of Aasen's ordered collections, which are the basis for his grammars and dictionaries, have been published and run into ten volumes (Aarset et al. 1992-2000). Aasen was supported and funded throughout, first by the Royal Norwegian Society of Sciences and Letters (DKNVS), later by the Norwegian parliament. Samuel Johnson and Ivar Aasen were both mother tongue speakers and self-taught mother tongue linguists. Both had their results accepted as a truthful and correct description of the mother tongue by their linguistic communities, and by the learned world[5] (which is not to say that those results were equally well liked by all). This immediate acceptance of results is closely allied to the mother tongue status of the people who did the work – they were part of the community where their languages belonged, and had been selected for the task because they were thought trustworthy.

3 Language documentation – who and how

Many languages are initially described in an attempted early standardised form through another standard language (Zgusta 1971: 304 f.). Ivar Aasen wrote his grammars in Danish and used Danish as the defining language in his dictionaries (Grønvik 1990: 247). Clement Doke (1931) and M. Hannan (1959 and 1974) did the same for Shona, using English as the language of description and the defining language.[6] But once the (semi-)standardised language becomes a school subject to children for whom the language is also the mother tongue, use of the mother tongue is required throughout, to make teaching efficient at all levels. Both definitions, explanations and meta language must be in the mother tongue. It is this part – defining, explaining, creating and establishing mother tongue meta language – which only a mother tongue linguist can do with the necessary professional authority.

This means that there are several qualifications required to become an efficient documentarist of language. Being a mother tongue speaker is one. Equally important qualifications are a genuine interest in language

4 "Til Slutning maa jeg da takke alle dem, som have ydet mig Hjælp og Bistand ved den Undersøgelse, som denne Bog er grundet paa" ('Finally I offer my thanks to all those who have given their help and assistance in the investigations which form the basis of this book').

5 Munch 1848 and 1850

6 In his classification of bilingual dictionaries, Zgusta (p. 304 and 306-7) points out "purpose" as a central category. When the purpose of a dictionary is to establish an as yet unrecognised (and not fully elaborated) standard language, linking it to a well-established language through a bilingual dictionary is a much used technique.

as such, one's own language in particular, and a thorough understanding of language as a system. The language documentarist should be a linguist. These days the training in linguistics is most easily obtained through formal education, and a degree in language studies or theoretical linguistics will always be an advantage. But Norwegian linguistic history shows that a true amateur – a person who loves the subject but is without formal training – can do a lot on his or her own, with some professional support[7], and the interaction between amateurs and academics is a constant feature in the history of Norwegian language collections (Skard 1932: 1 f; Grønvik 1997: 26). Some of the largest and most informative dialect dictionaries of Norwegian are produced by or based on collections from people with little formal education, and Norway has a wealth of dialect dictionaries.[8] First-hand knowledge of the language to be documented, combined with understanding and mastery of the functions of linguistic techniques, procedures and (current) best practice, is of more importance in achieving good results than a degree in language studies per se.

Since language is a common good to a community, and documenting language is a slow and labour intensive business, it is essential to have institutional guarantees that the documentation results will be cared for and kept in a state which facilitates new use. What is certain, is that close relations with a university, and a university library, is of great importance, both in relation to research, recruitment and productivity. A language documentation unit, tasked with maintaining linguistic standards (for orthography etc.) in a community, will also need to cooperate closely with the education authorities, and with any professional environment wanting to develop its mother tongue terminology.[9]

4 The documentation process

The process of establishing an orthography on the basis of speech was by Aasen's time fairly well agreed on in the Nordic countries. The favoured system was a phonemic transcription of synchronic speech, with some (variable) reference to older conventions and etymological considerations (cf. Hovdhaugen 2000: 888 f). There was in practice agreement among the learned that the system – the grid – must be described first, and the detail – the lexicon - must be collected and organised in accordance with the grid. Aasen's work plan of grammar first, dictionary next was considered and accepted by the board of DKNVS (Venås 1996 : 94 ff.). Today, the

7 The Norwegian Language Archives for Nynorsk at the University of Oslo are largely a product of voluntary work from 1930 onwards, cf. Skard 1932 p. 1 f. The Norwegian Dialect Archives (1936-2005) assisted a at a number of dialect dictionary projects.

8 Examples of from the multitude of dialect dictionaries are Paulsen 1981, Granlund 2008, Sandvik 1986, Sørensen 2004.

9 Language collections and documentation units are at the moment in a precarious position in many countries. For Norway, see http://sprakradet.no/Toppmeny/Aktuelt/Anbefalinger-om-en-samlet-ordbokpolitikk/. Danish dictionaries have lost their state financing, see http://sproget.dk/nyheder/carlsbergfondet-afvaergrer-kulturel-katastrofe.

same outline is set by every pioneer description of poorly documented languages, cf. the linguistics publications by SIL.[10]

What must be in place for a spoken language before a written standard can be set, is

- a choice of alphabet, and following on that,
- a description of the phonology and prosody,
- morphology (part of speech and inflection systems) and derivation,
- compounding and
- syntax

The reason for ensuring that these items are explicitly in place, and publicly recognised, is not properly part of the documentation process itself, but essential to the further use of documentation results, and therefore integral to the purpose of undertaking the documentation process. Once the written standard is set, in the form of grammars and dictionaries, it will be taught, quite likely as a school subject, and become a subject of examinations. Exams are important, so before a language becomes a school subject, the rules for how to write the language need to be established as reasonably clear, coherent and sensible. To neglect this issue can mean decades of delay, while the proposals for an academically documented standard language gather dust or become the victims of controversy and indecision.

Documenting languages that exist as non-standardised vernaculars has to be a process resembling a spiral, with many returns to former stages. The method can briefly be described in the following steps (a) collect a sample of materials, (b) analyse it and write out the analysis in the standard categorisation, (c) collect another small sample and repeat the process, comparing findings from the first sample with those for the second, and so on. All work must be based on verifiable materials, which are analysed, compared to a new sample, reanalysed, and of course regularly submitted to outside criticism and review, for instance by a reference group. The dependence on verifiable language materials and processing circularity are the most striking features of this method (Grønvik 2011).

The fairly standardised method can be operationalised as follows by a linguist who understands the language to be documented:

10 http://www.sil.org/resources/publications/publing; http://www.sil.org/resources/publications/ewp

1. Talk to a native speaker of the chosen language (which ideally is also the mother tongue of the interviewer)
2. tape a small sample (f.i. 15 minutes of continuous conversation) and store it safely (digital sound file)
3. Transcribe it provisionally at once (within twenty-four hours)
4. Segment it, analyse it and tag it as far as one can, leaving unanswered questions unanswered (never assume)
5. Collect another sample and repeat the analysis process
6. Compile lists of elements for different categories (phones/phonemes, morphs/morphemes, identifiable parts of speech, individual word forms, etc.)
7. Count occurrences of different elements and start sorting function elements from content elements
8. Draft a model lexicographic entry for different types of words and test it on your material
9. Repeat and refine (ad infinitum, or for as long as the funding allows)
10. At some point, compare several samples and start distinguishing a phonetic transcription from a possible common standard[11]

All of the steps in this procedure require linguistic insight. The most important linguistic insight is that of knowing the language as a mother tongue speaker – of understanding immediately what is said, and which word forms are intended. The procedure also requires personal qualities like a bent for analysis, a good ear, insight, patience and a will to work hard. Some training in practical linguistics is a prerequisite, and university studies in the selected language or language group is an great advantage, as mentioned above - the procedure has nevertheless been carried out by plenty of autodidacts, results ranging from brilliant to amateurish.[12]

Let me emphasise the following: Language documentation requires text from mother tongue speakers to base the analysis on, oral or written, published or unpublished. There has to be something to start from. Documenting a language in the 21st century should therefore always involve corpus building. For the major languages of the world, the text corpora to distill the analysis of the language from, have accumulated for centuries. For undocumented vernaculars, the corpus building has to happen hand in hand with the language analysis, and if there is nothing written down, taped and transcribed speech must form the basis of the corpus.

11 If there was available a good handbook for field linguistics, it would be referred to here, instead of setting the process out in detail. At present no such work is published.

12 In time even amateurish efforts can acquire value, witness some of the 18th century word lists from Norwegian dialects which are now incorporated in the Norwegian language collections.

5 The role of information and communication technology in documenting languages

The great advantage of computers over humans is evident in the handling of large quantities of data. Speed, stringency in classification and accuracy in dealing with large quantities of data all favour the use of computers in lexicography. The data can be stored, sorted, resorted, shared and manipulated with a speed and an accuracy no human handlers can match. The results of sorting can be saved and recombined. The speed and accuracy of computers in carrying out repeated operations on large quantities of data is what turns modern lexicographic projects into viable propositions in the first place, and it is a reasonable assumption that access to electronic language management has driven the revival of empirical language studies in the last decades.[13]

The ICT contribution to linguistics and lexicography in the area of subject analysis concerned with identifying categories, organising the relationship between them, and through this organisation of data, revealing new facts about the language. Every item gets classified, and researchers are therefore forced to deal with the poorly documented grey areas (for instance the borderline between appellatives and propria) and turn it into (documented) routine, instead of commenting in general terms and leaving the subject alone.

The next advantage of computers has to do with consistency. Every language consist of an unlimited number of items that share certain qualities, so that language-specific inventories of phonemes, morphemes, prosodic features and parts of speech can be extracted, and items can be grouped and named. All words have some sort of meaning which becomes apparent in context, and usages are often specific to a situation or a certain type of utterance or text genre. The linguist – not the computer – has to analyse a given language and discover its system, arrive at sets of categories and describe the relationship between them. But maintaining the chosen categories and the relationship between them, in describing each item consistently again and again throughout a long text, is impossible for a human being. This became more than evident in the 1980s, when several of the large, prestigious academic dictionaries off European languages were encoded and analysed. The process of digitisation revealed large quantities of minor slips and inconsistencies in all of them. The same phenomenon came to light when the first manuscript version, encoded in fields but without pre-set menus, of the *Duramazvi reChiShona* was proofread in January 1996.

13 See Zgusta 1971: 354 f. and Grønvik 2005.

The handling of language on a large scale by computers requires software (for corpora, dictionary writing systems, parsers etc) embodying agreed and set conventions, and a level of standardisation that had never been attempted before computers came into use. For lexicography and language documentation in general, this development has been very useful, because it forces researchers to make better and more explicit plans for how to inventory and organise the category system of their chosen language. Even if commercial software is used, adaptation and calibration will be necessary, and guidelines in the form of manuals and style guides reflecting best practice will be required at a very detailed level (Atkins and Rundell 2011: 118 f). In the planning process, the input of the linguist and the ICT specialist with a commitment to language study is of equal importance; no major study of language can be undertaken today unless both groups of professionals are involved.

Guidelines must reflect the linguistic and practical needs of the project and the language in question; to work out guidelines is a joint responsibility for mother tongue linguists and informaticians. A language with significant use of tone may need to mark the tone of lexical items. A language without a case system should not have a software format that makes filling in case schemas obligatory, as was the practice in school grammars of European languages with grammatical categories slavishly copied from Latin (Hals 1833: 11).

In terms of standardisation and operationalisation the use of computers in linguistics projects and language documentation has brought these research fields closer to the natural sciences than they were before. The research possibilities following on the use of computers have also brought about a renaissance of empirical studies and project within linguistics and lexicography.

In the work on poorly documented languages, the use of computers is an invaluable aid. Computerised language collections as well as dictionary databases with dictionary writing systems are available to all editors simultaneously, immediately after encoding. The increase in efficiency, compated to the pre-digital working environment, is enormous. This applies to all research projects, but has particular significance in dealing with pioneering research subjects with large quantities of data and low or dubious social status, which sadly often is the case when it comes to documenting languages.

6 Multilingual glossaries – pros and cons

Occasionally language documentation projects crop up where the organisers wish to skip the material collecting stage and get straight into dictionary production. These projects tend to be bilingual or multilingual,

and rest on the assumption that one can use a gloss list or a list of concept definitions in one of the world languages as a pivot for all other languages to connect to and through. Such efforts may provide useful first input towards machine translation, or make suggestive bases for multilingual word nets[14]. An older effort, aiming at coordinating European languages through mapping expressions for concepts, is the Atlas Linguarum Europae.[15]

What projects of this kind fail to do, is to create a reliable basis for the documentation of a vernacular as a stage on the way to developing a standard language. Gloss lists from different languages, matched against each other, and without a basis in running text, cannot disambiguate homophones or separate the senses of polysemes. A world register of concept definitions (in English)may work (just) for object description ("chair"), but will be bound to fail in describing transferred or metaphorical senses, or all the abstract and culture bound concepts embodied in any language. A multilingual gloss list or match against definition project is therefore no viable alternative to building language collections (in the form of corpora) as a basis for language standardisation.

7 Popularising the proposed standard language

Although my subject is language documentation, it is necessary to say something about how to gain social acceptance for a newly documented standard language, as the purpose of the documentation is bound to influence the form of the standard language itself. Relevant areas are

- choice of transcription system (alphabet) – which one is known to people already?
- sign inventory – special signs, digraphs or other signs for phonemes that lack a one to one correspondent in the chosen alphabet?
- level of purism – how should imported phonemes or word forms be handled?

Establishing a new standard language (based on the vernacular), or introducing it into new domains, is a big change in a language community, and should therefore be very carefully prepared. It is essential to move forward step by step and to allow space and time for some individual choice; the proposed change should be seen as an offer of increased possibilities and more freedom to language users who

14 Cf. for instance the mission statement of Universal Networking language (UNL): "The mission of the UNL Programme is to develop and promote a multilingual communication platform/infrastructure, with the purpose of enabling all peoples to share information and knowledge in their native languages" http://www.undl.org/index.php?option=com_content&view=article&id=47&Itemid=66&lang=en

15 Cf. http://www.kotus.fi/index.phtml?l=en&s=335

primarily have an instrumental interest in language. The term "language marketing" har arisen in the last few decades to describe the planning process of launching a standard (minority) language (Baker and Prys-Jones 1998: 221 f.)

In order to start teaching a new standard language in school, the following should (ideally) be in place: (1) the standard language itself, expressed in an (authorised and available) grammar and dictionary; (2) other teaching materials showing the language in use; (3) (preferably obligatory) training for teachers in using the new language and language tools in instruction; (4) agreement and/or permission from all involved institutions from Ministry down to each school (and school council); (5) information to and ideally (general) acceptance from parents[16]; (6) a plan for following and evaluating results in the initial period.

Underestimating the process of establishing the standard language as a viable alternative can have very unfortunate results. There are sad examples of total to partial failure where political initiatives to promote a local language as a language of instruction or administration have failed to prepare for expectations on the ground; one such case is the story of "malgachization" in Madagascar (Metz 1994 on language). People expect education to bring their children social advancement. Therefore, teachers as well as community must be prepared, teaching materials and examination systems ready, and social acceptance in place. In short – if the language users do not believe that changing the language of instruction is going to bring them personal and societal advantage, they will resist change, and opt for another language and culture instead, often that of former colonial masters.

8 Gaining acceptance

The effort of gaining acceptance for language documentation, as a step towards establishing a standard language, must to some extent be spearheaded by the mother tongue linguists themselves. They have to manage the collecting and guarantee the result; therefore, they have to motivate students, persuade funders and convince officials and school authorities (1) that their project is a good one, (2) that they are the right people to carry it out, (3) that those who assist them, will benefit somehow from project results.[17]

Even in the initial documentation process, quick and solid results are very important. People must see that things are happening. A small project, well executed, with results presented in a form everyone can see, will inspire

16 In Norway, adults can vote locally on the preferred school language, Nynorsk or Bokmål (Vikør 1993: 96)

17 Public acknowledgement of assistance may be enough. In the dictionaries produced by the ALLEX Project (1991-2006), all project contributors and facilitators are mentioned by name in the Introduction, (Chimhundu 1996: vi-viii.)

more confidence than a big project which takes long to materialise. By the same token, a grammar or dictionary which is useful to a large section of society should have early priority. The first ALLEX dictionary, Duramazwi reChiShona, took roughly 40 man years and five years to produce. A dictionary by two people working for 20 years would not have had the same impact.

Since a language documentation project today will be digital, it is also possible to give (limited) access to the language collections themselves from early on. This may be important in discussions on suggested rules for grammar, lemma selection etc. The collections prove existence, and thereby take away a prime cause for mistrust. The ALLEX Shona dictionaries were criticised for including too much rural vocabulary, claimed to be out of use long ago. Ir was vital to be able to refer critics to the Shona Corpus, collected during the project period, and accessible online.

All of this means that the pioneer mother tongue linguists to some extent also have to become politicians on the ground in the area of language and culture. Since documenting a language is such a massive economic and social investment, those in charge of the process – and the results – need to be known and trusted. A successful mother tongue documentation effort will of necessity involve a certain amount of outreach, and support for local efforts. In a linguistic environment with many unmapped dialects or minority languages, and uncertainty concerning linguistic identification, it is important to recruit knowledgable amateurs who can contribute information, and find young people who may be interested in a university degree in languages. The records of the Norwegian Language collections contain details of several hundreds of volunteers who have collected, transcribed, read and given feedback for up to 60-70 years.[18] Similar informants' records will be found in connection with all mother tongue archives in the Nordic countries.

One result of gaining acceptance for a proposed mother tongue standard is that other people start writing, publishing and experimenting with the written standard. This is a proof of acceptance, and therefore welcome, but the mother tongue linguist(s) may still dislike individual experiments, and have good reasons for doing so. Ivar Aasen was in constant correspondence on standardisation issues with a large circle of Nynorsk[19] supporters throughout his professional life, advising and correcting; but he was not always listened to (Venås 1996: 367 f.).

18 The digital slip archive for Nynorsk (Setelarkivet)(http://www.edd.uio.no/perl/search/search.cgi?tabid=436&appid=8), which lists the informant on each slip

19 The written standard Nynorsk was initially termed Landsmaal.

9 Mother tongue linguists or foreign field linguists?

Most of the world's languages could benefit from better (electronic) direct documentation, which today means speech and text in corpora. The corpus documentation of English is in a class by itself, and even for English one can see scope for improvement, for instance in the documentation of regional speech[20].

The languages needing documentation in order to prove their potential as standard languages, tend to be minority languages, or languages in a social position resembling that of minority languages, such as the native languages in in former colonies where the colonial languages remain as languages of administration and instruction. Non-standard languages are rarely taught as foreign languages in other countries (since they are not well described in the first place), so the foreigner has to learn them in situ. The undocumented languages of the former colonies which were described in the 20th century, were most often documented and described by people who learnt the languages on the spot as foreign languages. Some of them were self-taught linguists, others were professionals with heavy academic qualifications for their chosen task. The output, in the form of grammars, glossaries, dictionaries, text collections and meta literature varies enormously in quality, accessibility and scope.

If the point is to achieve rapid, comprehensive and exact documentation of the non-standardised vernaculars of the world, which strategy is best: training mother tongue linguists, or training more field linguists to come in from outside?

The only people who know the poorly documented languages of the world really well, are the mother tongue speakers. A linguist who is also a mother tongue speaker therefore has a much better starting point for documenting the language than a foreign field linguist who has to start with learning the language, the culture and a new way of life.
The additional advantages of the mother tongue linguist is that he - or she

- knows the essential concept register and the culture already
- has easier access to data and informants
- can more easily filter linguistic theory against the realities of the language and the culture
- can more easily obtain local institutional support

The advantage to the language community of having their language documented by a the mother tongue linguist is that he - or she (probably)

20 Speech corpora are very small compared to standard language text corpora, specialised dialect speech corpora tend to be even smaller. Cf. description of SED at http://sounds.bl.uk/accents-and-dialects/survey-of-english-dialects and of FRED (ca 1.1 mill words) at http://www.helsinki.fi/varieng/CoRD/corpora/FRED/background.html.

- lives in or near the area where the language is spoken, so that the language community can hope for continuous support
- is motivated not only by a wish to investigate an exciting language, but also understands the need for helping to create language tools for the community

- often is attached to a teaching institution in or near the community, and therefore can help with training students and finding ways of conserving and developing the field data.

The chief disadvantage that a mother tongue linguist faces, is the tendency to take his language and culture for granted, and therefore not examining it, or describing it, as extensively as is desirable. A researcher's detachment in relation to the mother tongue as an object of study can be hard to achieve. A mother tongue speaker can rarely inventory his own knowledge of the mother tongue, the way a learner of a foreign language can inventory their knowledge of the essentially unfamiliar language and culture that is the object of mastery.
The chief disadvantage of the non-mother tongue linguist is that he must learn the language and understand the culture to be documented, a task which can take years. The process of trust-building and adaption is also bound to be lengthy, while time to stay in the community may be short. Although a cultural outsider, the foreign linguist may nevertheless have some important advantages over the mother tongue linguist. These could be

- experience in documenting languages
- better access to getting results published
- a more thorough formal and practical training as a field linguist
- attachment to an institution with better conditions for funding work, and better support services
- better immunity to local pressures and interests

In my view, none of the advantages of the foreign field linguist outweigh having the language to be documented as a mother tongue. People commonly underestimate the amount of learning that goes into mastering a mother tongue, because we all have to do it. But mastering a mother tongue is a never-ending, long term process, and very few of us get to master two languages equally well.

10 The ALLEX Project – what did we do right?

For 20 years, 1992-2011, the UiO, GU and UZ were involved in a cooperative project to document and produce language tools for the African languages of Zimbabwe, later also for some of the African languages of Mozambique and South Africa.[21] This cooperation was highly successful and productive (cf. ALLEX Project website: ALLEX history and Chabata this volume). The cooperation produced (1) multi-use language collections (three corpora, two morphological parsers based on the corpora), (2) language products for public use (ten general and special language dictionaries), (3) a number of other publications), (4) ten completed Ph.D.s and (5) two language research centres.[22] For the first ten years GU was a partner[23].

This cooperation was a resounding success with far-reaching implications for especially two of the partners, UZ and UiO. Since many linguistics projects remain little known and never have much impact, it is important to look at the chief success factors in this cooperation. What follows here is a summary of the cooperation and the framework round it. A chief outcome – experiences from training students in the skills of field linguist work – will be dealt with in section 11 and 12 below.

(1) The original project initiative came from Zimbabwe. The UZ Department of African Languages had formulated a proposal and got support for it at all levels at UZ. They were therefore well prepared for hosting the cooperation, giving support and claiming ownership to results.

(2) A clear need was identified by the researchers behind the initiative – to produce the first monolingual general dictionary of Shona, and after that, monolingual dictionaries for other African languages of Zimbabwe. The UZ researchers also knew that UiO and GU had lexicography as an academic discipline. The first project (to be completed 1991-1995) spearheaded a research plan (in phases) covering twenty years and easily expandable beyond, which in practical terms meant that the UZ project participants were prepared for long term commitment. Since the Norwegian and Swedish project partners were institutional partners, their commitment for some time could also be taken for granted.

(3) The project was multi-professional and gave equal status to all academic fields – mother tongue linguists, lexicographers, computer scientists, and subject specialists (terminology). It is generally agreed today that large scale lexicography and linguistics projects are impossible

21 The ALLEX Project (1991 - 2006) and the CROBOL Project (2007-11).

22 African languages Research Institute (ALRI), UZ, University of Zimbabwe and Eduardo Mondlane university, Maputo

23 Språkdata, GU.

without tailored software and good ICT solutions. It follows that project teams should include ICT specialists as team members of equal standing with the linguists. Software tailoring involves problem-solving, and for creative joint problem-solving – expressing linguistic and lexcicographical concerns through programming – the ICT developers should be team members.[24]

(4) The team composition and division of labour was suited for the task. The UZ team were specialists in African languages and literature, and mother tongue speakers of the chief African languages of Zimbabwe. A few of them, amongst them the project leader, Herbert Chimhundu, were linguists. The Norwegian and Swedish participants encompassed mother tongue linguists, a corpus specialist, and information technologists. They knew about best practice and state of the art for corpora, language collections and lexicography, and categorisation and organising data in relational databases. They also knew that lexicographical research products require documentation, properly stored in language collections, since lexicography implies language standardisation, and the standardisation process can evoke controversy. Norwegian linguistic history in particular encompasses experience on how to process a language from the vernacular stage to a written standard (Haugen 1976: 405 f.).

(5) The project proved expandable. When the UZ partners realised that they would have to build oral language collections in parallel with working on the dictionaries, they were able to involve BA students, guided by UZ staff, in a massive field work effort resulting in corpora for the two languages Shona and Ndebele. Printed text was added as the corpora grew, in order to expand vocabulary.

(6) Project aims came first. In multinational projects, cultural boundaries and attitudes and other personnel issues can both aid, hinder and destroy project aims. The chief challenge is competence drainage – student and staff acquiring knowledge and skills and then leaving the project. To counteract this possibility, all training in Zimbabwe was made accessible to all participants from scholarship students upwards, irrespective of degree or status. Project procedures were documented, and no field was left to one person only. Some training (basic computer skills, knowledge of languages involved) was obligatory for project participants at UZ. Planning and budgeting was discussed in plenary meetings. This was all necessary in order to keep up focus in project work.

24 Cf. for instance Dragland 2014

(7) Funding was directed towards project aims and an important project aim was long term institutionalisation of lexicography and language collections at the universities in the South. This essentially meant spending on recruitment (guest researcherships, Ph.D. scholarships; staff conference participation in the South) and field work, and minimum expenditure on the Nordic participants (beyond institutional compensation for project participation). It also meant holding back on project travel for the senior project participants (conferences, field work participation etc.).
This funding policy made the project work attractive to recruits and institutions in the South, but also less popular with the administrations in the North (initial tight funding, and not much in overheads).[25]

11 How to train mother tongue linguists

When the ALLEX Project started, task number one was training a lot of people at the University of Zimbabwe in the skills a documentation linguist must have.

In order to get corpus materials for the planned dictionary, The ALLEX Project organised a huge collection of oral data across the country, using students and staff at the university of Zimbabwe, none of whom had done anything like it before. They had to learn interviewing, recording, and transcription in record time. The UZ project leadership set up a district list for the country, and got formal permission from the authorities to send out interview teams. Teams were organised so that students would go primarily to their home districts, to minimise linguistic and social distance in the interview situation. Topic lists were compiled and work routines established, staff members in the project at the UZ were tasked with following up each their group of student teams. Secretarial staff at UZ were taught encoding and tagging of transcribed interviews. The Nordic contribution in all this was concentrated around staff training in ensuring quality, choice of software, construction of databases and overall work plans.

One of the first items called for was a handbook for linguistic field work, which simply set out what to do if you wish to collect oral material and use it as data for scientific work - a sort of "best practice" summary for linguists. When I asked experienced field linguists, both within Norwegian mother tongue linguistics and linguists with interest further from home, they immediately started emphasising the difficulty of the task, the years of experience needed to get reliable results, leaving me with the impression that asking BA students at an African university to conduct a massive

25 This is a recurrent issue in connection with North-South research cooperation. Cf. f. i. the NUFU Annual report 1997: 13 "NUFU's restrictions on supporting salary and overhead costs make it very difficult for Norwegian Research Institutions to be involved in cooperation programmes without having a guarantee in advance for covering these costs from other financial sources than NUFU."

collection of oral data was a very risky procedure. Nor did they know of any elementary teaching materials. There was no such thing to be had, so interaction and hands-on learning became the order of the day.[26]

The project summing up of the oral data collection is twofold.

(1) It is possible to train mother tongue BA students to collect oral materials of adequate quality for corpora and other research purposes. The initial teaching does not need to be long, slow or heavily theoretical, because the main points and techniques are not difficult to understand or practice. The really important thing is project organisation and follow-up, and a project leadership that is prepared to pitch in and do any job in demand. Age, maturity and knowledge is a great advantage in the interview situation, but youth, courage and honest curiosity brought in some original contributions that the experienced among us would never have achieved.

(2) Both students and staff became proficient in data collection for mother tongue linguistics much faster through learning by doing than they would have done if we had required theory (and exams) first. The task of collecting, transcribing and assessing the quality of their own mother tongue data had the effect of driving students and staff in search of literature and theory to sort out issues of sound variation, derivation systems, compounding and sense development which they had to handle in order to meet production results. At the end, results showed up in exams. The students who had participated in data collection did better than those who had not; their experience contributed to the knowledge of their year as as whole, so everyone did better than before. This was noticed and commented upon by the external examiners.[27] The most motivated of the oral collections students in turn became MA students on scholarships and research assistants, some going on to Ph.D. studies and a lifelong commitment to mother tongue studies, linguistics and lexicography.

Another more general lesson from the ALLEX Project is that linguistics would have more motivated and confident practitioners if students were introduced to dealing with the raw data of language much earlier in their courses, allowed to do more, and to see results of their own contributions. In the course of the twentieth century mother tongue studies, as well as linguistics in general, have accumulated a considerable research and teaching literature, which is all to the good. But to give it meaning it has to be connected with first hand study of language. It

26 Today there are some university web sites, but the standard handbook for collecting and storing oral materials is still unwritten.

27 Comment from Professor Chimhundu at the time, when he held the position of Dean of the Arts Faculty at UZ.

would be good to see the craft aspect of linguistic documentation brought more to the fore, while research literature to a greater extent could take its proper place as reference literature. Documentation procedures should be taught to mother tongue students, to facilitate MA studies built on original materials collected by the students themselves.
Of the procedure outlined in section 4, Steps 1-3 were taught to the students taking part in the ALLEX Project corpus collections. Steps 4-7 are difficult both for mother tongue linguists and for field linguists with another language as their mother tongue, but the closer the linguist's own language is to the language to be described, the better the starting point is. A linguist who is a native speaker will know from his own understanding of the language where utterances (periods) stop, what order elements normally appear in, what constitutes a (complex) word, what words or phrases are used as names and what words belong to the general language. The mother tongue linguist will sort allophones and allomorphs and recognise homonyms and polysemes with far greater confidence than the linguist whose own linguistic background is remote from the language to be described(Fortune 1979:23 f.). Steps 8-10 is work that require composite experience and should ideally be done in a group under sound linguistic management.

12 General lessons

No developed country uses a foreign language which is not understood by the majority population as its written language of instruction and administration. All developed countries have sizable linguistic communities to whom the language(s) of instruction and administration represents the mother tongue. Over the past 3-400 years, modernisation has included institutionalising the care and development of written standard languages, in private academies, publishing houses or in public institutions.

To have one or more national standard languages which represent the mother tongues of the linguistic communities in a country is a mark of identity and a cause of national or community pride. But the most important thing about a developed standard language is that it is a powerful storehouse of the thought and imagination embodied in a culture. Standard languages can be the means of organising and developing a community consciousness which no modern society can do without, simply because they contain tens of thousands of words whose meaning and use the linguistic community agree on. This vast and detailed agreement means that every user of a standard language can speak and write of any topic and expect to be understood by the rest of his linguistic community. It means that the standard can be taught with confidence as a school subject, and used as a matter of course in public administration and in the management of daily life. To have the freedom of a written standard language essentially means that the language user can take the language

toolbox for granted, and concentrate on conveying what he or she wants to express as efficiently as possible. [28]

If there is no common, standardised written language in a community managed through writing, another written language must be used instead, and the linguistic community must put up with learning the other language in addition to their own languages. To have to use a foreign language in speech and writing in daily life entails a burden for the individual, and a loss of efficiency and speed for the community as a whole. To use a modern metaphor communities with written standards based on their mother tongues have access to broadband, while communities who transact their business in foreign standard languages, have to use phone boxes and dial their way through.

All infrastructure needs maintenance and development. Roads don't build themselves, they have to be planned as part of a road net, and once built, they require renewal and adaptation to new requirements. Organised communities have public agencies responsible for planning and maintaining transport, as a matter of policy. Organised communities also need to store their language products - that is what national libraries are for - and they need (preferably public) agencies to survey language development, decide on standardisation issues and provide language tools of various kinds. Today, this means maintaining, developing and expanding electronic language collections (corpora, full form generators, parsers etc.), and seeing to it that the population is equipped with the language tools and the language management each society needs to deal with its daily business.

13 Concluding remarks

In an ideal world, language needs should be recognised and development projects funded, wherever the need may be located. As things are, the source of funding language description in developing countries often lies in a developed country, and depends on an interest in remote languages at university level. The best that can be achieved is therefore very often some form of cooperative project where the funding comes from the North. If such projects however focus on the need in the developing country to achieve ownership to the documentation of local languages, there is a good chance that new cooperative projects will arise which are as good as the ALLEX and CROBOL Projects or even better.

The state of linguistics in Africa has changed a great deal over the last twenty years, especially in the training of mother tongue linguists. One particular fact illustrates this: for the first Ph.D. candidates from the

28 The points expressed here are dealt with as linguistic rights in UDLR 1996 Article 7 to 9.

ALLEX Projects, it was hard to find members from African countries for the examining board and evaluation committees. When the last CROBOL Ph.D. candidates reached the examination stage, former ALLEX participants helped man the examining boards.

In this article, the principles and practices of language documentation are set out as the author sees them. The formulation of these principles are drawn from a long and largely cooperative experience in mother tongue studies, language planning and language documentation. It follows that whatever is useful and well thought out here to a large extent is a collective product of best practice as seen by the working research and documentation teams the author has been a member of. This is as it should be. Language is a social product and language documentation gains credibility only when it represents community consensus as well as lexicographical best practice. Particular credit for shaping the best practice of the ALLEX Project from the start belongs to Herbert Chimhundu (UZ), Daniel Ridings (then GU) and Christian-Emil Smith Ore (UiO). Whatever is right in the pages above is shared with colleagues; whatever is wrong or misrepresented belongs to the author.

14 Literature

ALLEX Project website: *The Allex Method.* http://www.edd.uio.no/allex/allex%20method.html

Aarset, Terje; Nes, Oddvar; Bondevik, Jarle 1992-2000: *Skrifter frå Ivar Aasen-selskapet* / Serie A Tekster. Bergen. Ivar Aasen-selskapet. Band 1-10.

Aasen, Ivar 1873: *Norsk Ordbog med dansk forklaring.* Christiania. P. T. Mallings Boghandel.

Auroux, Sylvain et al. 2000: *History of the Language Sciences.* Vol. I p. 1-1094; vol. II p. 1095-2004. Berlin – New York. Walter de Gruyter.

Baker, Colin and Prys-Jones, Sylvia 1998: *Encyclopedia of Bilingualism and Bilingual Education.* Clevedon. Multilingual Matters.

Chimhundu, Herbert et al 1996: *Duramazwi reChiShona.* ALLEX Project. College Press. Harare.

Doke, Clement M. 1931: *Report on The Unification of The Shona Dialects.* Carried out under the auspices of the Government of Southern Rhodesia and the Carniege Corporation. Printed for the Government of southern Rhodesia. Hertford, England. Austin and sons ltd.

forskning.no: *Norske IT-forskere i verdenstoppen.* Ved Dragland, Åse Dragland (2014). http://www.forskning.no/artikler/2014/mars/386302

Fortune, George 1979: *Shona Lexicography.* In: Zambezia 1979. p. 21-47.

Granlund, Thorbjørn 2008: *Gamle ord og uttrykk fra Osen : skrevet på dialekt.* Osen.

Grønvik, Oddrun 1992: *The Earliest Dictionaries of Nynorsk in the Light of Present-Day Dictionary Typology.* In: Poulsen 1992. p. 247-258

Grønvik, Oddrun 1997: *Om kjeldegrunnlaget for Norsk Ordbok*, In: Vikør 1997 p. 23-38.

Grønvik, Oddrun 2005: *Norsk Ordbok 2014 from Manuscript to Database - Standard Gains and Growing Pains.* In: Papers in Computational Lexicography Complex 2005. Budapest: Linguistics Institute, Hungarian Academy of Science 2005 ISBN 963-9074-35-7. s. 60-70.

Grønvik, Oddrun 2011: *Lemma collection for the first monolingual dictionary.* In: Lexander et al. 2011. p. 269-278.

Hals, Ove 1833: *Uddrag af Modersmaalets Sproglære og Retskrivningslære : udarb. til Brug ved Underviisningen i den nederste Classe af de lærde Skoler . .* Christiania : Abelsted.

Hannan,M. 1959: Standard Shona Dictionary. Harare: The Literature Bureau.

Hannan, M. 1974: *Standard Shona Dictionary.* Harare: The College Press/The Literature Bureau. 1st ed. 1959.

Haugen, Einar 1976: *The Scandinavian Languages. An Introduction to their history.* London. Faber and Faber ltd.

Hovdhaugen, Even 2000: *Normative Studies in the Scandinavian Countries.* In: Auroux et al. p. 888-893.

Johnson, Samuel 1755: *Dictionary of the English Language.* J. & P. Knapton, London. I-II.

Lexander, Kristin Vold et al. 2011: *Pluralité des langues, pluralité des cultures* : Regards sur l'Afrique et au-delà. Mélanges offerts à Ingse Skattum à l'occasion de son 70ème anniversaire. Oslo. Novus Forlag.

Metz, Helen Chapin 1994: *Madagascar: A Country Study.* Washington: GPO for the Library of Congress.

Munch, Peter Andreas 1848: Anmeldelse. *Det norske Folkesprogs Grammatik af Ivar Aasen.* In: Myhren 1975. s. 24-38.

Munch, Peter Andreas 1850: Anmeldelse. *Ordbog over det norske Folkesprog, af Ivar Aasen* In: Myhren 1975. s. 39-49.

Myhren, Magne 1975: *Ei bok om Ivar Aasen.* Språkgranskaren og målreisaren. Oslo. Det Norske Samlaget.

Paulsen, Ragnhild 1981: *Ordbok over Nøttlandsmålet omkring 1900.* Tønsberg. Vestfold Historielag. XV,

Poulsen, Jóhan Hendrik W. et al. 1992: *The Nordic Languages and Modern Linguistics 7.* Proceedings of the Seventh International Conference of Nordic and General Linguistics in Tórshavn, 7-11 August 1989, Vol. I-II. Føroya Frodskaparfelag (Annales Societatis Scientiarum Færoensis, Supplementum XVIII), Tórshavn 1992

Sandvik, Sigurd 1986: *Gamle ord frå Suldal.* Samla av Sigurd Sandvik. Suldal : Suldal mållag.

Skard, Sigmund 1932: *Norsk Ordbok. Historie – plan – arbeidsskipnad.* Oslo. Det Norske Samlaget.

Sproget.dk: *Hjemmeside for sprogspørsmål.* Kulturministeriets institutioner for sprog og litteratur, Dansk Sprognævn og Det Danske Sprog- og Litteraturselskab.

Carlsbergfondet afværger 'kulturel katastrofe'. 29.10.2014. Sproget.dk/nyheder/carlsbergfondet-afvaergrer-kulturel-katastrofe.

Sørensen, Torstein 2004: *Ord fra det gamle Nesna.* Stamsund. Orkana Forlag og nordland fylkesbibliotek.

University of Zimbabwe. Department of African Languages and Literature: *Proposal to establish an African Languages Research Institute 16 February 1999.* http://www.edd.uio.no/allex/reports/ALRI_Institution_proposal_1999.pdf.

Venås, Kjell 1996: *Då tida var fullkomen.* Ivar Aasen. Oslo. Novus forlag.
Vikør, Lars 1997: *Norsk Ordbok – nynorskens leksikografiske kanon?* Rapport frå eit seminar på Blindern 31. mai 1996. Oslo. Universitetet I Oslo – Institutt for nordistikk og litteraturvitskap.

Vikør, Lars 1993: *The Nordic Languages. Their Status and Interrelations.* Nordic Language Secretariat. Novus Press.

Zambezia 1969-2005. *The Journal of Humanities of the University of Zimbabwe.* ISSN: 0379-0622.

Zgusta, Ladislaw et al. 1971: *Manual of Lexicography.* Janua linguarum,Series maior; 39. The Hague : Mouton.

Universal Declaration on Linguistic Rights (UDLR) 1996. World Conference on Linguistic Rights. Barcelona, Spain, 9 June 1996.

Universal Declaration of Human Rights (UDHR).1948. The UN General Assembly. Paris. 10.12.1948

URLs

ALLEX Project:
http://www.edd.uio.no/allex/
Ethnologue:
https://www.ethnologue.com
British Library:
http://www.bl.uk/learning/langlit/dic/johnson/1755johnsonsdictionary.html
Sproget.dk:
http://sproget.dk/
SIL:
http://www.sil.org/resources/publications/publing;
http://www.sil.org/resources/publications/ewp
UNL:
http://www.undl.org/
Country Studies US:
http://countrystudies.us/madagascar/
Nynorsk slip archive (Setelarkivet):
http://www.edd.uio.no/perl/search/search.cgi?tabid=436&appid=8
Språkrådet (The Norwegian Language Council):
http://sprakradet.no/Toppmeny/Aktuelt/Anbefalinger-om-en-samlet-ordbokpolitikk/.
Freiburg Corpus of English Dialects: (FRED-S):
http://www.helsinki.fi/varieng/CoRD/corpora/FRED/
SED:
http://sounds.bl.uk/accents-and-dialects/survey-of-english-dialects
UDHR:
http://www.un.org/en/documents/udhr/
UDLR:
http://www.unesco.org/cpp/uk/declarations/linguistic.pdf

Abbreviations

ALLEX = African Languages Lexical Project (1991 – 2006)
ALRI = African Languages Research Institute, University of Zimbabwe
CROBOL = Standardization and Harmonization of Cross-border Languages (2006 – 2011)
FRED = Freiburg English Dialect Corpus
GU = Göteborgs Universitet (University of Gothenburg)
LD = Language documentation
SED = Survey of British Dialects
SIL = Summer Institute of Linguistics
UDHR = Universal Declaration on Linguistic Rights
UDLR = Universal Declaration on Linguistic Rights
UN = United Nations
UNL = Universal Networking Language
UiO = Universitetet i Oslo (University of Oslo)
UZ = University of Zimbabwe

Summary

There are about 7000 languages in the world. Roughly 2000 of them have a written standard of sorts. About 100 are national languages in general use in their countries. A majority of the languages of the world will go largely undocumented for a long time unless a more efficient way is found of documenting them, as a first step towards establishing them as standard languages. The key to improving efficiency in language documentation is to train mother tongue speakers to be linguists with the necessary practical skills. This is a possible, but underutilised approach. In this paper, the issue of mother tongue documentation is explored in the light of experience gained in a North-South cooperation to document and in some cases create written standards for the African Languages of Zimbabwe, between the University of Zimbabwe (UZ), the University of Gothenburg (GU) and the University of Oslo (UiO) lasting from 1991 to 2011 to document the African languages of Zimbabwe, Mozambique and South-Africa.
The overall goal of the paper is to set out the principles and practices of language documentation (LD) in general, The Allex-project as an example within the field of LD, with discussion and concluding remarks). This is dealt with under the following points:
Why document languages (sections 1-2)
How to document languages (sections 3-9)
The Allex Project (sections 10-12)
Concluding remarks (section 13)