Bridging the Gap between Language Documentation and Description for Effective Linguistic Research

[1]Aniefon AKPAN, [1]Eno-Abasi URUA & [2]Moses EKPENYONG
[1]Department of Linguistics and Nigerian Languages; [2]Department of Computer Science
University of Uyo, Nigeria.
[1]daniels.mmefon@gmail.com; [1]eno-abasiurua@uniuyo.edu.ng;
[2]mosesekpenyong@uniuyo.edu.ng

Abstract

Language Documentation and Language Description have been treated as different fields within Linguistics over the years and this has widened the gap between these two fields even though they interrelate at the data and analysis levels. Language Documentation is a relatively new research area in Linguistics which focuses on multimedia development and digital archiving of endangered languages, while Language Description provides description and analysis of language data. This linguistic research gap has contributed to the retarded process of documentation, analysis and accountability of endangered languages in Nigeria. This paper therefore aims at proposing strategies that will bridge the gap that exists between these two fields. The proposed approach is primarily through collaboration between them and other related fields such as Computer Science. For effective collaboration, workshop training and seminars on Natural Language Processing Toolkit, descriptive and documentary theories and practices are organised to balance the information flow between these two fields. Collaboration between them should yield substantial language output for the language and linguistic research communities. Furthermore, the collaboration should address delay, funding and human resources involved in the processes of describing and documenting languages. Data for this study were obtained through fieldwork in the Ibibio community (Nigeria). This approach should ensure effective linguistic research in the 21st Century, where digital research is at the forefront.

1. Introduction

Language Documentation emerged as what is considered a sub-field of Linguistics in 1995 (Austin, 2014) and has received considerable attention from linguists all over the world. Its emergence over the last decade was a response to the rising need to record and account for languages all over the world, especially extinct and endangered languages, to ensure that the future generation of that speech community will have a reference material for their language and some form of linkage to their roots. Language Documentation is not

a completely new enterprise (Himmelmann, 2006); it has been practiced within Linguistics for some years now. The compilation of written historical and cultural speech events was the concern of historical linguists in the nineteenth century and this is one of the influences on Documentary Linguistics. Even though Historical Linguistics is one of the influences on Language Documentation, the approach and methodology of Language Documentation differs from the tradition of the philologists. As a new field, there are researches aimed at establishing standards for tools and models, goals and methods (Himmelmann 1998; 2006).

Twenty years after the emergence of Language Documentation, there are still separation issues between Language Documentation and Language Description. On the one hand, there are proposals to keep the two activities separate and on the other hand, to treat them as one. Language Description or descriptive linguistics as the name implies, is concerned with providing the morphological, syntactic and phonological description of a language, while Language Documentation is concerned with making records of speech events of a community available to a wide range of users. (Himmelmann, 2006). Language Documentation is an activity that was assumed under Language Description. There are recently developed misconceptions concerning Language Documentation which Himmelmann (2012:187) has outlined, including the following:

1. Language Documentation is concerned with collecting heaps of data;
2. Language Documentation is primarily concerned with technology and digital archiving;
3. Language Documentation is not concerned with analysis mainly because of its concern with technology and digital archiving;
4. the assumption that it is not analytic in nature.

These misconceptions about Language Documentation have led to the belief that it is a non-linguistic enterprise. Language Documentation aims at providing a comprehensive language data that reflects the social and cultural contexts of a speech community (Himmelmann, 1998), but these linguistic systems and sub-systems that have been recorded will be incomprehensible without analysis.

Apart from the misconceptions about Language Documentation, there are challenges facing it, especially in the African region. In Africa, Language Documentation has met numerous challenges that have hindered the documentation of most African languages. Urua (2013) outlined these challenges to include: number of languages in the continent, lack of documentary skills and research funding. In addition, Wilbur (2014) points out that technical infrastructure, inexperience and prioritizing of language material make the task of documenting local languages in Africa difficult.

Apart from the challenges observed by Urua (2013) and Wilbur (2014) in documenting African languages, there is the unacceptability of Language Documentation as a linguistic venture by the mainstream linguists in most African regions. The separatist

issue between Language Documentation and Language Description in addition to the misconceptions about Language Documentation have created a gap between the two subfields, which has further delayed the documentation of most endangered languages in the African region. Hence, this paper aims at presenting a framework to bridge the gap between Language Documentation and Language Description to accelerate the documentation of languages in Nigeria (and other African Regions), as well as enhance linguistic research in the 21st Century.

2. Similarities and Differences between Language Documentation and Description

Language Description has been functional in the description of little known languages, previously unrecorded languages and well documented languages and it comprises two activities. The process of describing the linguistic features of a language is divided into two: high-level and low-level. Himmelmann (1998) views the high level as the collection, transcription and translation of primary data and the low level as the analysis of the data. The first activity is considered as the documentary activity which involves the representation of the recorded language data and the second activity as the descriptive activity. Himmelmann (1998) argues that these activities should be kept separate even though they are closely interrelated for epistemological, methodological and practical reasons.

These two activities overlap primarily in the aspect of data. In every field in linguistics, just as in most other disciplines, data is the foundation on which most researches are built. The two activities require data collection for analysis. The aim of Language Documentation is to provide comprehensive records of a wide range of speech events in a particular speech community. There are diverse speech events in a speech community and it would be of immense benefit and value to the speech community, humanities and related fields if the comprehensive record of each speech activity in the community is provided. It is however not vivid as to what degree or depth the comprehensiveness of the record of a language should be. It is in this regard that Austin (2008) made the claim that the notion of what makes up a comprehensive record of a language is still unclear. By contrast, Language Description aims at producing (mostly) a collection of texts and grammars.

Differences between these two activities can be observed in their goals, research methods, and areas of interest (see Austin and Grenoble, 2007; Himmelmann, 1998). Language Documentation employs sampling, reliability and naturalness in its methodology. The methodological issues involved in Language Description include justification of analysis, and definition of terms and levels. In addition, these two activities also differ in their workflow. Audio recording serves as the source for the creation of analytical representations in the form of lists, summaries and analysis in Language

Description. In Language Documentation, audio or video recording are used to create representation for archiving or mobilization, (Austin, 2014:63).

To further distinguish Language Documentation from Language Description, Himmelmann (2006:15) outlines the following crucial features of Language Documentation:

1. Focus on primary data: The aim of Language Documentation is to collect and analyse diverse range of genre of data that can be made available to large groups of users.
2. Explicit concern for accountability: Availability of primary data makes the evaluation of the analysed linguistic data possible. Hence, transparency is implied in the provision of data from the unprocessed data to the processed data.
3. Concern for long term storage and preservation of primary data: Language Documentation concerns itself with the quality of recordings and metadata to ensure the location and evaluation of the documented resources by users in the future.
4. Work in interdisciplinary teams: Achieving comprehensive Language Documentation requires expertise and contribution of knowledge from diverse disciplines in addition to the fundamental linguistic expertise needed for annotation.
5. Close cooperation with and direct involvement of speech community: Language Documentation involves the native speakers of the community in the documentation project as co-researchers who determine the overall output of the documentation. The activities involved in the documentation require little academic training which provides room for the native speakers to be involved in the project.

Language Documentation focuses mostly in the speech community. Its interest is getting the language community involved in the documentation of their language. In order to achieve this aim, the members of the speech community are trained on the necessary tools required for the documentation. The tools range from video camera, audio recorders to software. In addition, aspects of the processed data or output of the documentation is given back to the community. It may be in form of orthography (in the case where the speech community does not have any orthography), alphabet chart, picture books, and recorded language data on DVD, etc. This is one of the outstanding practices observed in Language Documentation that is not obvious in Language Description.

3. The Gap between the Two Activities

The argument to keep Language Documentation separate from Language Description has gone on for decades now since the conception of Language Documentation. Scholars like Himmelmann, Austin and Wilbur have tried to provide the line demarcating Language Documentation from Language Description in theory. There

may be a seemingly reasonable explanation for the on-going researches by linguists towards providing a line separating these two activities. In view of the above, (Austin, 2008) questioned the meaningfulness of the separation and the right place to draw the line between these two activities. The gap created in the process of separating these two activities can be observed in:

1. Primary data: Data are essential aspects of research. It is the base on which evaluations and conclusions are derived. Linguistics makes use of data for analyses and evaluation of language properties. The primary data employed in Language Documentation differs from Language Description with regards to the volume of data obtained from the fieldwork. Language Documentation aims at providing records of speech events in a community, hence it aims to gather large amounts of data that will enable efficient and effective analysis and processing of the data. The bottom line is that the two activities require data for analysis although they differ in the quantity required.
2. Analytic Approach: The analytic approach adopted in Language Documentation differs from Language Description. Language Documentation's analytic approach is computationally-based while the analytic approach adopted in Language Description tilts towards the traditional linguistic mode of analysis. Language Documentation is involved with building language databases using language software and employing algorithms or software for the analysis.

We chose to bring up the above two points because they are paramount in any linguistic documentation or research. The gap created by these two outstanding activities has gone beyond the activities to the responsibility of linguist in accounting and recording the languages in the world.

It appears that globally, Language Documentation has not actually gained its place in linguistics and it is also viewed as a non-scientific enterprise (Himmelmann 2012), because Language Documentation is considered to be concerned primarily with technology and digital archiving. Considering the workload and responsibilities increasingly channelled to linguists due to the number of languages going into extinction or endangered yearly, it calls for collaborative research within linguistics. We are not proposing the merging of these two activities but rather collaboration to quicken documentation of languages in Nigeria and strengthening of linguistic researches in Nigeria in the 21st century.

4. Approach

The workflow of Language Documentation and description observed in (2) above, will be considered here but from a collaborative perspective. On the one hand, the approach will be concerned with the task of documenting and accounting for languages in Nigeria,

and on the other hand, it will look at linguistic research as a whole in Nigeria in the 21st century.

In Africa, there are multiple languages in the region, and a hand full of these languages, are yet to be described and documented. According to Bamgbose (2006), over 2000 or 30% of nearly 6,700 languages in the world are African languages. Prah (2003) notes that 85% of Africans speak between 12 to 15 African languages. These languages may be a cluster of mutually intelligible speech forms or dialects of the same language.

There are numerous languages in Nigeria but the total number of languages varies among linguistic authors. The range in which the total number of Nigerian languages are fixed fall between 100 to 527 and this has been one of the setbacks encountered in the documentation of Nigerian languages. In an attempt to address the linguistic situation in Nigeria, a few surveys and researches have been performed in Nigeria towards determining the total number of languages. For instance, Coleman (1958) puts the number at 250, Tiffen (1968) puts it at 150, Bamgbose (1992) stretches the number to 400, Gital (1998) provides a range between 400-500 and in Ethnologue edited by Simons and Fennig (2017), it is put at 572. The criteria for determining a Nigerian language varies from one author to another, hence the variation of the total number of Nigerian language. In addition, there is the problem of where a dialect ends and a language begins, which has caused the number of Nigerian to increase and decrease between authors and the scenario in which a dialect is termed a language by its custodian because they do not want to be associated with a dialect for psychological, social and political reasons, makes the duties of the linguists difficult.

Documentation of Nigerian languages as used in this context relates to documentation in the sense of Language Documentation sub-field of the linguistic discipline. Documentation assists in determining the number of languages in a region and also aids in accounting for languages (active, moribund, extinct).

Simons and Fennig (2017), report in the Ethnologue that there are about 7 extinct and 44 dying languages out of the 527 languages in Nigeria. In the South-South Geopolitical region of Nigeria, many languages or varieties have no orthography and written text, they are not used in the media (prints, electronic, social etc.); they are not taught in schools, etc. (Udoh, 2003). Despite the conflicting reports in the number of (extinct) languages in Nigeria, the implication of the linguistic situation in Nigeria calls for urgent documentation of the languages in the country. The workload inherent in documenting Nigerian is too voluminous to be left to only documentary linguists in Nigeria considering that there are only few of them versed with the knowledge of the art and techniques of Language Documentation. Hence, in view of the above, we propose the following approach to quicken the documentation of Nigerian languages.

1. Collaboration: To achieve what Himmelmann (2006) terms a comprehensive record of a language (active or moribund), there is need for collaboration between

Language Documentation and other disciplines or activities within linguistics. Although, it may not be possible to determine qualitatively and quantitatively, the point in which a research or documentation of a language is complete (Austin, 2014:62), it is necessary to ensure that the core linguistic aspect of a language and the highly sensitive areas/topics (taboos, belief, tradition etc) of a language and culture are covered. Collaboration between Language Documentation and Language Description will ensure the analyses and documentation of more varieties than the two activities operating separately and moreover, the resources (human and finance) required for fieldwork will be reasonably reduced. In documenting a language or analyzing a linguistic aspect of a language, a field work is required to gather the data. The three stages involved in fieldwork are: Pre-fieldwork, Fieldwork, Post-fieldwork which involves time, effort, finance and personnel. The instrument, (video, audio etc.) employed in obtaining data in the field are necessary expenditures towards the overall objective of documentation and research. The workload involved in each activity – Language Documentation and description, is intensive considering the resources involved.

This workload involved in the two activities will decrease because audio and video records made in Language Documentation projects of a language in addition to the field notes taken, can be subjected to further analyses in descriptive linguistics (phonology, morphology, syntax etc). Moreover, there is need for collaboration between computer science and linguistic (all activities). This would further quicken the documentation of Nigerian languages. This will be achieved through programming (algorithm that will be used to process primary data). The programming will be done at the second stage of the documentation after collection of data and representation of the recordings. Also, collaboration between geography and linguistics (Language Documentation) will help produce language maps (electronic and print) for the languages in Nigeria.

2. Training: Although collaborative research is crucial in the 21st century, it is important that linguists should be trained with the state of the art linguistic applications which are used for analyses. These application tools reduces the workload inherent in analysing primary data, such applications include parsers, Praat, syntactic complexity analyser, etc. Most of the analyzers are built to work with English data such as the syntactic complexity analyzer (Lu, 2010) but could be modified to work on Nigerian languages through collaboration with computer scientists. There are numerous Natural Language processing (NLP) tools available for analysing Natural Languages which should be utilized in processing and documenting Nigerian Languages. In the 21st century, where a good number of researches revolve around computer packages, there is need for linguists to have a fair knowledge of the computer packages relating to linguistic analyses. This approach of computational analyses reduces workload and saves time in analyses and production of research materials. The traditional method of hard paper write-

up prior to transference to a word processing pack, consumes time and delays the processing and production of research materials. To achieve this, training on the software tools for Nigerian linguists is a necessity and should be a requirement.

3. Training Workshops: There is need for training workshops on documentation and conservation of Nigerian languages and cultures. This will ensure that the issues encountered in the process of documenting Nigerian languages are addressed, and possible solutions proffered to raise the progress level in the documentation projects.
4. Output: Given the fact that descriptive linguistics is viewed as an ancillary function to documentary linguistics and vice versa (Himmelmann, 2012:204), the expected preliminary output to be derived from documentation projects (that is the representation of the wide range of language use in a community) is expected to serve as input for descriptive linguists to produce grammars, dictionaries and collection of texts. Austin (2014:63) righty captures it when he summarises that documentary linguistics and descriptive linguistics are complementary activities with complementary outcomes. In addition, Austin and Grenoble (2007) point out that linguistic analysis reveals speech genre, lexical forms, and grammatical paradigm or under-representation in a documentary record. The output of documentary language projects in Nigeria is indeed an input for analyses in descriptive linguistics which grammars and dictionaries can be produced. The output from the implementation of an algorithm in processing language data can be used by linguists for analyses and description in different linguistic sub-fields. This collaboration will help reduce the time spent in annotation and analyses of language data by documentary linguists.

5. Collaborative Framework between Language Documentation and Description

Language Documentation seeks to collect and analyze properties of languages using linguistic theories and processing tools. Conversely, descriptive linguistics provides the rules governing written languages in semantics, syntax, morphology, phonology, etc. This section presents a framework in which Language Documentation and description can collaborate to reduce time spent in data collection in a language and the analysis of the different linguistic aspects of the language. The input for the collaborative framework is speech events in a given speech community. The speech events in a speech community are recorded in the form of audio or video materials. Prior to the main recordings, the metadata for the documentation would have been obtained. The recorded speech events serve as data for both documentary and descriptive linguistics from where analyses can be made.

This collaboration reduces resources spent during fieldwork because instead of employing different resources (human, equipment) for the data collection in the two activities, one channel of resource is used. The recorded multimedia will be represented in the form of transcription, translation or annotation and archived in digital and non-digital

forms, from which analyses in descriptive and documentary linguistics can be made. The multimedia database will also serve as input for descriptive linguistic analyses and other related fields. From the representations made; morphological, phonological and syntactic analyses can be performed. In addition, morphological segmentation, part-of-speech tagging, parsing, word segmentation, speech recognition, etc. are possible analyses which can be performed in the field of documentary linguistics from the archived data. The target audience for Language Description and documentation include the language community, linguists, pedagogy and other disciplines. The outputs from Language Documentation analyses are functional for further researches in related fields of linguistics.

This collaboration between the two activities within linguistics will strengthen linguistic research in Nigeria and also quicken documentation and description of languages in Nigeria. The collaboration framework of the two activities is captured below:

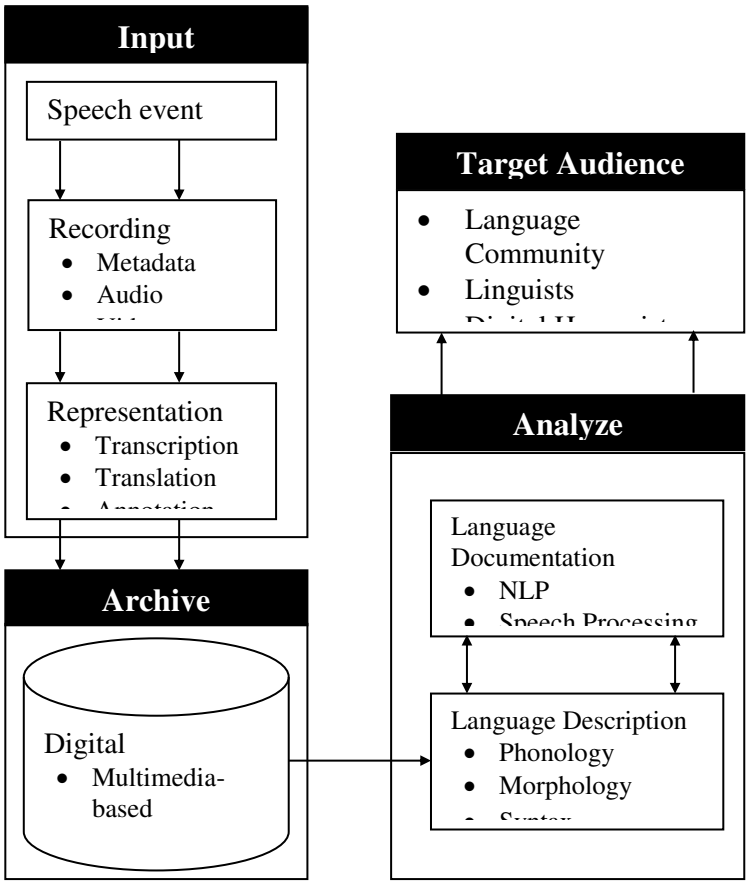

Figure 1. Collaborative framework between Language Documentation and description

6. Implementation of the Collaborative Framework

The input needed for the above proposed collaborative framework is language data which acts as the foundation of every significant research. The language data for this research was obtained from the Ibibio speech community, where there are diverse speech events. The above collaborative framework between Language Documentation and Language Description was developed by Akpan, Ekpenyong and Urua (2017) was applied to the research done on "Developing an audio annotator for the efficient documentation of Ibibio children's play songs/games" to test the efficiency of the framework. The speech event used for the research is the Ibibio children's play songs/games in the Ibibio community. The recorded Ibibio children's play songs/games are performed mostly in the Ibibio language – a regional language spoken in Akwa Ibom State, although there are instances in which the Ibibio children engaged in plays using the Efik language, a related and highly mutually intelligible form of Ibibio. The time which the children get together to play is normally during the evening hours at home, or during break time and game time in school. The Ibibio children's play songs/games were recorded in audio and video formats, in addition to the metadata and field notes taken. The video format is vital because it portrayed the Ibibio cultural context, climate, facial expressions, gestures, etc.

These video contexts aided in the interpretation of the data. Using the recorded data, descriptive notes were made on the sociolinguistic, phonological, morphological and syntactic aspects observed in the Ibibio children's play songs/games. The resources that would have been employed in gathering another set of data for the descriptive analysis of the diction of the Ibibio children's play songs were preserved and directed towards other language projects because the data employed for the documentation aspect were also used in the descriptive aspect, hence bridging the gap of performing a documentation of the Ibibio children's play songs/games in the documentation sense by a linguist and performing a phonological or morphological context of the Ibibio children's play songs from another set of data by another linguist.

Furthermore, the representations of the recorded visual and audio Ibibio children's play songs/games in the forms of transcription, translation and annotation provided a base for further research and also serves as a database in digital and non-digital formats where reference could be made for further research. The play songs/games were produced in a compact disc and given back to the Ibibio speech community in line with Language Documentation ethics; and the database created would serve as input for researches in linguistics and other digital humanities. The representation of the data was processed using the ELAN (Eudico Language Annotator) tool.

The target audience for the research included the Ibibio language community, linguists and digital humanists.

7. Conclusion

Language Documentation is a new development in the field of linguistics with diverse opportunities for research and training. This activity (also known as documentary

linguistics) focuses on the collection, analysis and mobilisation of recorded speech events with the aim of preserving and disseminating the processed data to a wide range of users including the speech communities. There have been arguments to keep Language Documentation separate from Language Description for practical reasons. Conversely, the two activities have been viewed as complementary especially at the output level. An outstanding feature of Language Documentation is its collaborative nature which is proposed in this work between Language Documentation and Language Description and other related fields. The collaborative framework presented in this work aims at speeding up the documentation of Nigerian languages. Using the Ibibio language, the framework was tested to assess its efficiency and effectiveness. The collaboration addresses delay, funding and human resources involved in the processes of describing and documenting languages because when these two activities join forces, the volume of resources employed for the fieldwork and analyses should be reduced, all things being equal. Thus, this collaboration should aid in increasing the number of documented and archived languages in Nigeria and in addition, enhance linguistic research in Nigeria.

References

Austin, P. 2008. Language Documentation and Language Description. Paper presented at 3L Summer school.

Austin, P. 2014. Language Documentation in the 21st Century. *Journal LIPP*. 3. 57-71.

Austin, P. & Grenoble, L. 2007. Current trends in Language Documentation. *Language Documentation and Description*. Volume 4: 12-25. London: SOAS.

Bamgbose, A. 1992. *Speaking in Tongues: Implications of Multilingualism for language policy in Nigeria*. Ibadan: Wemilse Press.

Bamgbose, A. 2011. African languages Today: The challenge of and prospects for empowerment under globalization. In: Bokamba, E., Shosted, R. & Ayalew, B. *Selected proceedings of the 40th annual conference on African linguistics.* Somerville, MA: Cascadilla Proceedings project.

Coleman, J. S. 1958. *Nigeria: Background to Nationalism.* Berkeley: University of California Press.

Gital, M. M. 1998. How many Nigerian Languages. In: Arohunmolase, O. (Ed.). *Nigerian Languages for National Development and Unity.* Ibadan: Lolyem communications.

Grimes, B. 2000. *Ethnologue*, volume 1. Dallas: SIL International.

Himmelmann, N. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161-195

Himmelmann, N. 2006. Language Documentation. What is it and what is it good for? In: Gippert, J., Himmelmann, N. & Mosel, U. *The essentials of Language Documentation*. New York: Morton de Gruyter.

Himmelmann, N. 2012. Linguistic data types and the interface between Language Documentation and description. *Language Documentation and conservation*. Vol 6. 187-207

Lu, X. *2010*. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15.4: 474-496.

Simons, G. F. and Fennig, C. (Ed.) 2017. *Ethnologue: Languages of the World, 20th Edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com

Tiffen, B. W. 1968. Language Eductaion in Commonwealth Africa. In: Dakin, J. Ed. *Language in Education: the problem in commonwealth Africa and the Indo-Pakistan Sub-continent*. London.

Udoh, I. 2003. *The languages of the south-south zone of Nigeria: A geopolitical profile*. Lagos: Concept Publication.

Urua, E. 2013. Different Climes, Different Bird Songs: a peep into Nigeria's linguistic tapestry. Paper presented at the Nigerian Academy of Letters Publishers, Lagos.

Wilbur, J. 2014. Archiving for the Community: Engaging Local Archives in Language Documentation projects. In: Nathan, D. and Austin P. Eds. *Language Documentation and Description, Special Issue on Language Documentation and Archiving*. 12: 85-1