AN ACTIVITY BASED SPOKEN LANGUAGE CORPUS OF NEPALI

Jens Allwood[1], Bhim Narayan Regmi[2], Sagun Dhakhwa[2], and Ram Kisun Uranw[2]

[1]University of Gothenburg, Sweden
[2]Centre for Communication and Development Studies, Nepal

jens.allwood@ling.gu.se, rite2sagun@gmail.com, bhimregmi@gmail.com, rkuranw@yahoo.com

ABSTRACT

Language is used for communication and communication facilitates social activities. If we want to capture this, linguistic investigation has to be carried out within a wider context. Examination of linguistic communication in a wider context shows that it is multimodal. In order to study naturalistic multimodal communication using a corpus, the corpus should contain a combination of recordings, documentation, and transcription of multimodal communication from different social activities in naturalistic settings, preserving unedited conversation. This paper presents a brief account of the principles, methodology, current status, and preliminary findings, based on an incrementally growing and multimodal activity based spoken language corpus of Nepali.

Index Terms— activity based, multimodal, spoken language corpus, NSC, Nepali language

1. INTRODUCTION

The Nepali Spoken language Corpus (NSC) is being developed (with funding support from Swedish Research Council (VR) and SIDA), in collaboration between the University of Gothenburg, Sweden and the Centre for Communication and Development Studies in Nepal. It is aimed at continuing and completing the work on a spoken language corpus carried out under the NELRALEC project [1] [2], and at analyzing some central features of spoken Nepali. The current goal is the collection of a 500 K words corpus with pauses, silences and overlaps annotated, where our purpose is to analyze features of spoken language, like pauses, silences, overlaps, feedback, and own communication management.

Nepali has 11 million native speakers in Nepal according to the national census 2001 [3]. There are two other Nepali corpora – one spoken language corpus, and another written language corpus. Both of the corpora are genre based [4] [5] [6] [7]. The corpus presented here is social activity based and, thus, differs theoretically as well as methodologically from the other spoken language corpus.

Taking the concept of "social activity related communication" as a basis for the study of language and communication is not new, the seeds were planted as early as 1953, in the concept of "language game" suggested by Wittgenstein, and later developed as a concept by Allwood in 1976, 2000, and 2007 respectively [8] [9] [10] [11]. The conceptual goal of a social activity based language and communication study is to describe, understand and explain linguistic interaction, especially face-to-face, direct, multimodal communication and the factors that condition such interaction, see Allwood [12].

Though the methodology of such a study is open and can include different methods, the main type of method is, as Allwood (pp. 1-2) has noted, recordings, registration and analysis of authentic linguistic interaction, in as "non-arranged", "naturalistic" circumstances as possible with a primary focus on face-to-face, direct, multimodal communication. However, communication using different kinds of communication technology, such as telephones or computers can also be taken into consideration [13]. This approach has already been applied to the spoken language corpus at Gothenburg University and is followed in the Nepali corpus too [14].

2. KEY POINTS IN DEVELOPING NSC

Some of the key points around which the corpus has been developed are presented in the following subsections.

2.1. Social activities

There are many activities in society where people communicate using language, gestures, etc. The constitutive features of these social activities affect both the verbal and gestural communication in the activities. Thus, a corpus based on social activities can be used for the study of many different aspects of human communication such as language and gestures, sounds and silences, individual contributions and interactive patterns etc.

2.2. Interaction

Use of language in a social activity is often interactive, though there are activities that are more monological, such as lectures, TV and radio broadcasts or expressive uses of language in artistic and ritual functions. In our corpus, the focus is on interaction.

2.3. Naturalistic settings

Though we do not deny the importance of the data from controlled settings/studio environments, e.g. in extracting speech features for speech modeling, naturalistic settings provide a wider perspective on human communication. Since communication is not limited to features of phonemes or prosody, but goes beyond to topic, emotions, social status, activities, and combinations with gestural means of communication, this naturally led us to considering naturalistic settings as primary. The corpus also includes some radio talk shows and TV talk shows produced in a high quality studio environment and with controlled behavior where the naturalness may be questioned. However, a radio talk show and a TV talk show can also be regarded as social activities, where the studio environment and controlled behavior of the participants is natural to this social activity, so it is also a naturalistic setting.

2.4. Unedited media

Human communication has many repetitions, incomplete verbal units, long pauses and silences, overlaps, and variations in pitch, loudness, length, etc. which can only be captured in unedited audio and audio-video recordings. Thus, in this corpus most of the recordings have not been edited, but we have not been able to prevent editing in the radio talk shows and TV talk shows.

2.5. Contributions

The basic units of interactive talk are contributions. They are often vocal verbal or gestural and can be characterized as a participant's communication from the point where it begins to the point where another participant makes her/his contribution or interrupts. Contributions include utterances, silences, and gestures, and are part of a broader concept of communication where verbal and gestural elements can function together. In this corpus, contributions are the starting point for an analysis either of relations between contributions or of elements within contributions.

2.6. Multimodality

Communication is realized through various related simultaneous moods, such as speaking-hearing, gesturing-seeing, writing-reading, etc., we could call this audio, audio-visual, visual and written modes. This corpus aims at multimodality, where at least speech is transcribed in the case of audio data, and speech, gesture, and visual form of the setting are annotated in the case of video data.

2.7. Transcription

Transcription includes the rendering in written form of contributions containing utterances, pauses, silences, overlaps and gestures. If used this way, the term "transcription" has an overlapping meaning with the term "annotation", but the two terms can also be separated so that "transcription" means written recording of speech and "annotation" is slightly wider, including transcription of the linguistic units and information about the linguistic functions, gestures and context etc.

2.8. Metadata

Every transcription and recording is accompanied by metadata. The metadata covers information about the speakers and recorded activity. They also include social and geographical information about the speakers and the participants, setting, necessary artifacts and duration of the recorded activity. See below, sections 4.5 and 4.6.

2.9. Ethical issues

We have obtained prior informed consent from all participants to be recorded in the Nepali spoken language corpus.

3. SOCIAL ACTIVITIES IN NSC

We started with a list of 18 social activities based on Swedish Spoken Language Corpus at Gothenburg University. As we aimed to carry out comparative studies between Swedish and Nepali spoken language, we did not modify the list, but instead left some of the social activities empty and added some new activities, more typical of Nepal, to the list. Currently, there are 24 activities in the list, but it may become longer as new activities are identified. The list is as follows: (1) Shopping, (2) Discussion, (3) Court proceedings, (4) Task oriented formal meeting, (5) Dinner conversation, (6) Family gathering, (7) Conversation while working (weaving, farming, etc.), (8) Quarrel, (9) Hotel, (10) Academic seminar, (11) Radio talk show, (12) TV talk show, (13) Interview, (14) Hospital, (15) Classroom interaction, (16) Story telling , (17) Phone, (18) Market Place, (19) Task oriented informal meeting, (20) Honor, (21) Fortune telling, (22) Formal discussion, (23) Thesis defense, (24) Elicitation.

Currently there are 220 recordings with a total duration of 61:35:33 in the NSC (Table 1). In the data, 133 files (with a duration of 40:58:15) have been transcribed and contain 386314 words. The words in the 87 untranscribed files (with a duration of 20:37:18) have not been counted. However, we estimate that when they are counted, the total number will be larger than the target 500 K words.

Table 1. Current situation of NSC

Activity Title	Activity code	Duration	No of recorded Activities
1. Shopping	1	01:17:01	6
2. Discussion	2	06:46:20	29
3. Court proceedings	-	-	-
4. Task oriented formal meeting	4	01:31:46	2
5. Dinner Conversation	5	00:58:01	3
6. Family gathering	-	-	-
7. Conversation While working	7	02:42:31	17
8. Quarrel	-	-	-
9. Hotel	9	00:07:49	1
10. Academic seminar	-	-	-
11. Radio talk show	11	09:17:17	20
12. TV talk show	12	01:25:16	2
13. Interview	13	04:14:43	7
14. Hospital	14	03:24:12	33
15. Classroom interaction	15	00:41:14	3
16. Story telling	-	-	-
17. Phone	17	02:11:37	14
18. Market place	18	00:23:34	2
19. Task oriented informal meeting	19	02:52:34	9
20. Honour	20	02:14:54	7
21. Fortune telling	21	02:32:03	6
22. Formal discussion	22	05:11:43	13
23. Thesis Defence	23	01:24:10	5
24. Elicitation	24	12:18:48	41
	Total	**61:35:33**	**220**

Among the empty activities listed above, activity type no (3) Court proceedings is not possible to record in Nepal because of lack of legal provision to permit recording of such activities. The activities of (6) Family gathering, (8) Quarrel, and (10) Academic Seminar have not yet been recorded. Activity (15) Classroom interaction has 3 recordings and has been transcribed recently but they are still to be annotated for pauses, silences, overlaps, and activity (16) Story telling has just been recorded but not yet transcribed. So the NSC is an incrementally growing corpus with the aim of getting as many social activities collected as possible.

3.1. A brief description of the social activities

A brief description of the social activities in NSC is presented below. There are 133 files that have been annotated for pauses, silences and overlaps, but, as we have already mentioned, 87 files are yet to be coded for these features. The words in the transribed 133 files have been counted and are presented with the duration and number of words in Table 2.

3.1.1. Shopping

Shopping is the activity of visiting shops to buy goods. In this activity a customer visits a shop, asks a shopkeeper about the quality, size, price, manufacturer and durability of the goods, bargains at the end and buys if he is satisfied.

3.1.2. Discussion

Discussion is the activity of talking about a topic. In this corpus the term is used for open talk between two or more people about any topic. Most of the casual conversations are grouped under this activity.

3.1.3. Court Proceedings

A court has well established procedures for filing cases, petitions, summons, advocating, and judging and finally ordering implementation of the law. “Court proceedings” refers to such procedures. Unfortunately, because of the lack of legal provision for providing us with permission to record, we have not had the opportunity to record any activity in this category.

3.1.4. Task oriented formal meetings

This is the kind of meeting where the topic of discussion is pre-defined and the procedure is formal. A task oriented formal meeting either leads to a decision on a certain task or ends at some point on the way to a decision.

3.1.5. Dinner conversation

Dinner conversation is talking while having dinner. However, We have grouped also lunch under this activity. The activity perhaps could be better represented with the term “meal conversation”.

3.1.6. Family gatherings

Here we have in mind a gathering of family members where there may be close family members or close relatives. We have not had the opportunity to record this activity. In many cases, such gatherings are related to private matters such as property, relationship between family members, etc., where recording might be sensitive.

3.1.7. Conversation while working

This activity type involves talking and working simultaneously. Talk itself is a kind of social work, but here work is taken in a more narrow sense to mean accomplishment of a physical task such as weaving, weeding, sewing, etc.

3.1.8. Quarrel

This is fighting-with-words, different from debate or discussion in that it mostly involves the emotion anger. A quarrel mostly contains words or structures that are normally supposed to be offensive, abusive, and hurtful and sometimes it results in fighting as well. It cannot easily be arranged and since it usually happens suddenly, it is difficult to get this activity recorded. Even if we get an opportunity to do so, there is a risk of physical attack by the persons involved in the quarrel. We do not have any recordings of this activity so far.

3.1.9. Hotel

This activity is related to talk between hotel personnel and owner or guests concerning facilities, number of rooms, rates, reservations, etc. There is only one recording of this activity in NSC.

3.1.10. Academic Seminar

This is an academic activity where academicians meet and talk on a topic in a formal way. Because of the many participants and many speakers from different sides, it has been difficult to record. There is no recording of this activity yet in NSC.

3.1.11. Radio Talk Show

A radio talk show is a radio program where a journalist invites a person/s to discuss or express their views on a given topic. During the conversation the journalist leads the invitee/s to a topic or intervenes with queries, feedback, etc. It is a semi controlled conversation recorded in high quality studio settings which is useful when we want to extract speech features for language processing purposes.

3.1.12. TV Talk show

A TV talk show is similar to a radio talk show, only differing in media in terms of quality, however, there is a vast difference in terms of modality. It is audio-visual, while radio is only audio.

3.1.13. Interview

An interview is talk about a person's experiences, ideas, views, etc., e.g. involving a journalist or researcher and an interviewee to make the interviewee's thought, experiences, life, etc. known to others.

3.1.14. Hospital

In this activity, the hospital personnel may be doctor, health worker or administrative staff, talking to the patients or the patient's caretakers.

3.1.15. Classroom interactions

This involves teacher and students' talk concentrated on teaching-learning activities.

3.1.16. Story telling

This is the activity of telling a story where a story teller and listeners take part. Story telling is a very well known folklore activity, practiced in the villages of Nepal. It has a special setting, style of conversation and special communicative roles for the participants. Unfortunately, it is disappearing gradually. We have a few recordings of these activities that are yet to be transcribed.

Table 2. Duration and number of words in some of the activities in NSC

Activity Title and code	Duration	No of activities	No of words
1. Shopping (1)	01:05:12	4	11893
2. Discussion (2)	03:57:21	16	45257
3. Task oriented formal meeting (4)	00:29:53	1	4495
4. Dinner Conversation (5)	00:58:01	3	9202
5. Conversation While working (7)	01:51:27	5	16727
6. Hotel (9)	00:07:49	1	1346
7. Radio talk show (11)	09:17:17	20	90735
8. TV talk show (12)	01:25:16	2	15978
9. Interview (13)	04:14:43	7	38811
10. Hospital (14)	03:24:12	33	23230
11. Phone (17)	02:11:37	14	22664
12. Market place (18)	00:19:04	1	2112
13. Task Oriented Informal Meeting (19)	01:54:58	3	17816
14. Honour (20)	02:14:54	7	16181
15. Fortune telling (21)	02:32:03	6	21789
16. Formal discussion (22)	03:30:18	5	33020
17. Thesis Defence (23)	01:24:10	5	15058
	40:58:15	**133**	**386314**

3.1.17. Phone

This is telephone communication. It is not face-to-face but direct and has its own patterns, styles, and vocabulary.

3.1.18. Market place

This is an open market activity where many buyers and sellers are involved in activities such as promoting/advertising their goods, bargaining, choosing, buying and selling, etc.

3.1.19. Task oriented informal meeting

This is a task oriented meeting without formal procedures.

3.1.20. Honor

This is an activity of felicitation where a person is awarded a kind of recognition letter, praised, and some other people make a speech on his/her contribution to society, biography, etc. Such meetings are growing in popularity in Nepal.

3.1.21. Fortune telling

This is an activity of talking about a person's past, present, and future by a professional fortune-teller.

3.1.22. Formal discussion

This is a discussion carried out in a formal way.

3.1.23. Thesis defense

This is an academic activity where a student presents his research findings orally and the experts make queries in order to evaluate his work.

3.1.24. Elicitation

This activity involves question-answer during field research where a researcher makes queries to a person who knows about a topic and the person answers. In our corpus, it differs from interviews in that it is specific to a certain topic and involves a researcher, who tries to get specific information and often has a wider perspective than is common in interviews.

4. METHODOLOGICAL CONSIDERATIONS

The methodology followed in building the NSC is presented below in brief.

4.1. Selection and environment of recording

As social activities are the basis of the corpus, natural settings have been chosen for the recordings. Thus, most of the recordings have been made in their actual environments. For example, a recording of a weeding takes place in a nursery garden and includes the background noise produced there and a recording of a dinner conversation includes noises produced with tools and utensils. There are also studio recordings, since radio talk shows and TV talk shows are also important social activities, which are normally recorded in a studio environment.

4.2. Tools for recording

We have used a Sony handy cam and mp3 digital recorder, and Samsung’s mp3 digital recorder in most of the cases. But the data from radio and TV have been recorded with their own equipment which we got from their archives. The data has been recorded af a 44khz frequency rate.

4.3. Sample size of the recordings

There is no fixed digital size or size in terms of duration or number of words for the recordings. Since it is based on naturalistic social activities, the size is determined by the activity. For example, a social activity entitled ‘hospital’ includes doctor-patient conversation and the size of that recording is determined by the time given by the doctor to that particular patient as in the recording V001014023 which is a 1 minute and 10 seconds long video recording containing 221 words. Similarly, there is no predictable time and number of words ratio. For example a recording of an interview A001013001 which has a duration of 00:31:17 contains 4739 words whereas another interview V001013003 with longer duration of 00:39:22 (more than 8 minutes longer) contains 4378 words (only 361 less words).

4.4. Naming conventions

The NSC naming conventions have been established in the following way: There is a letter A or V for audio and video respectively followed by nine numbers grouped into each 3 sets for corpus label, activity label and activity number. For example, V001002003 stands for the ‘third’ (003) ‘video recording’ (V) of the social activity ‘discussion’ (002) within the Nepali spoken language corpus, a part of the Nepali National Corpus developed in the NELRALEC project’ (001). Likewise A002024001 is the ‘first’ (001) ‘audio recording’ (A) of the social activity ‘elicitation’ (024) within the ‘Nepali spoken language corpus developed under the Nepali and Lohorung Spoken Language project’ (002). This convention allows us to have 999 recordings within 999 social activities, within 999 corpora as its maximum.

4.5. Speaker information

The following information (metadata) concerning participants is noted while recording and maintained in the corpus: (1) Name, (2) Age, (3) Gender, (4) Mother tongue, (5) Second language, (6) Dialect (social and geographical), (7) Education, (8) Place of primary education, (9) Profession, (10) Address (including contact phone number and email). This information can be helpful to carry out sociolinguistic, dialectal and second language studies using the corpus.

4.6. Information on recorded activities

The following information (metadata) is maintained in the corpus: (1) Transcription status, (2) Recorded activity ID, (3) Recorded activity title, (4) Short name, (5) Recorded activity date, (6) Tape, (7) Anonymity, (8) Access, (9) Activity Type (at three levels), (10) Activity purpose, (11) Activity roles, (12) Activity procedures, (13) Activity environment, (14) Activity artifacts, (15) Duration, (16) Participants, (17) Recorder, (18) Transcription name, (19) Transcriber, (20) Transcriber's ID, (21) Transcription date, (22) Transcribed segments, (23) Transcription system, (24) Checker, (25) Checking date, (26) Description, (27) Comment, (28) Start and end times, and (29) Section.

4.7. Transcription and annotation

Nepali is written in Devanagari script with its long tradition of writing. However, standardized writing has become different from speech in the course of time. Thus, in order to maintain closeness to spoken language, the speech of the recordings have been transcribed phonemically in Devanagari Unicode.

The format has been based on the Gothenburg Transcription Standard (GTS), where spoken language features such as pauses, silences, overlaps, unclear speech, broken words etc. are annotated [15].

5. RESEARCH BASED ON NSC

Some preliminary research has been carried out based on the Nepali spoken language corpus. There is a paper on the intonation patterns of Nepali feedback units and another paper on the relation between writing and speech and the functions of a multifunction Nepali word *chaahin* [16] [17]. These papers present only preliminary results, but show the importance of research in the field of spoken language, multimodal communication and corpus based studies and the validity of such a corpus for studies of Nepali.

Although not yet formulated as research papers, there are many interesting features of spoken language and communication that can be found while working with the corpus. To note a very few, spoken Nepali is different from written language not only in pronunciation but also in terms of vocabulary and grammatical structure. There is even a set of words and a set of functions, which are found only in the spoken language. Another interesting feature of interpersonal communication connected to overlap and pause is that most of the overlaps follow pauses in the conversation. However, these and other issues need to be supported by quantitative analysis and appropriate explanations drawing on the research to be carried out.

10. SUMMARY AND CONCLUSIONS

Summarizing the above discussion, the Nepali spoken language corpus is being developed as a continuously growing corpus, which is now ready for research on spoken language features and other communicative features. The notion of social activity is the basis of the corpus and the corpus is multimodal, based on unedited recordings in natural settings, containing detailed information about participants and the recorded activity itself. We hope it will provide new opportunities of research and a wider perspective on language and communication for researchers.

10. REFERENCES

[1] http://bhashasanchar.org/ncorpus_spoken.php

[2] [7] Yadava, Y. P., A. Hardie, R. R. Lohani, B. N. Regmi, S. Gurung, A. Gurung, T. McEnery, J. Allwood and P. Hall, "Construction and annotation of a corpus of contemporary Nepali", *Corpora* Vol. 3 (2), Edinburg University Press, UK, 2008, pp. 213–225.

[3] Central Bureau of Statistics National Planning Commission Secretariat His Majesty's Government of Nepal (CBS) and UNFPA, "Population Census 2001: National Report", CBS and UNFPA, Kathmandu, 2002.

[4] http://cqpweb.lancs.ac.uk/bandhu/index.php?thisQ=corpus Metadata&uT=y

[5] http://cqpweb.lancs.ac.uk/nncv2/index.php?thisQ=corpus Metadata&uT=y

[6] http://bhashasanchar.org/ncorpus_written.php

[8] Wittgenstein, L., *Philosophical Investigations*. Oxford, Blackwell, 1953.

[9] Allwood, J, "Linguistic Communication as Action and Cooperation", *Gothenburg Monographs in Linguistics 2*, University of Göteborg, Dept of Linguistics. 1976.

[10] Allwood, J., "An Activity Based Approach to Pragmatics", In Bunt, H., & Black, B. (Eds.) *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. Amsterdam, John Benjamins, 2000, pp. 47-80.

[11] [12] [13] Allwood, J., "Activity Based Studies of Linguistic Interaction", *Gothenburg Papers in Theoretical Linguistics 93*, Department of Linguistics, Göteborg University, 2007.

[14] Allwood, J., "The Swedish Spoken Language Corpus at Göteborg University", *Proceedings Fonetik 99:The Swedish Phonetics Conf. June 1999,* Göteborg University, Sweden, 1999.

[15] Nivre, J., J. Allwood, L. Grönqvist, M. Gunnarsson, E. Ahlsén, H. Vappula, J. Hagman, S. Larsson, S. Sofkova, and C. Ottesjö, *Göteborg Transcription Standard v6.4.*: Department of Linguistics, Göteborg University, 2004.

[16] Allwood, J. and B. N. Regmi, "Intonation Patterns in Nepali Feedback Units", *Proceedings of Oriental COCOSDA 2010,* Paper 61, November 24-25, Kathmandu, 2010.

[17] Regmi, B. N., "*ChaahiN*: a case study in terms of variants, frequency and function", *Proceedings of Oriental COCOSDA 2009*, Poster 10, Aug.10th -12th, Beijing, China, 2009.