Corpus-based analysis based on Bodding's Santal Dictionary

Minegishi, Makoto[*] Takashima, Jun[†] Murmu, Ganesh[‡]

26-28 November 2007

Paper Presented at the third International Conference of Austroasiatic Linguistics

1 Introduction

Santali, the most populous member of the North Munda subgroup of the Munda family, is spoken in eastern states in India. The language is said to have at least two dialectal groups; southern and northern. The main phonemic difference between these dialects is number of vowels: the southern dialect has six vowels, whereas the northern dialect has eight or nine vowels, though dialectal studies of the language seem yet to be elaborated and dialectal boundaries have not been clearly shown so far.

Minegishi & Murmu (2001) is an example of southern dialects, as it has a six vowel system, /i, e, a, o, u, ə/, the last of which seems to have been derived from /a/. The description is based on the pronunciation of the second author, who was born in East Singhbhum District of Jharkhand State,i.e, southern part of former Bihar state.

Rev. P. O. Bodding's Santal Dictionary (1932-1936), hereafter called as 'BSD', represents northern dialects, as it has eight vowel system /i, e, e̲, a̤, a, o̲, o, u/. BSD describes the front vowel /e/ as having "several values, the mid-front-narrow (like in Norw. fred), the mid-front-wide e (like in Engl. men), or the mid-mixed-narrow (or wide) e", and "/e̲/, the low-front-narrow or low-front-wide sound, pronounced like the vowel in Engl. air or dead".

As to the back vowels, as Bodding describes , "/o/ is the mid-back-narrow-round or the mid-back-wide-round vowel sound, something like the sound in "note." The lips are not much protruded", and "/o̲/ is the low-back-narrow-round, the low-mixed-narrow, or the low-back-wide round sound, long or short, like in Engl. law, or not".

Despite phonetic varieties, according to the above description, basic vowel system of BSD in modern phonemic symbols would be /i, e, ɛ, ə, a, ɔ, o, u/.

[*]Research Institute for the Languages and Cultures of Asia and Africa (ILCAA), Tokyo University of Foreign Studies, Tokyo, Japan

[†]ILCAA, Tokyo

[‡]Ranchi University, Ranchi, India

BSD is a monumental work, providing not only descriptive linguistic data, but also a rich reference of cultural information of Santali society. In 1990's, as part of Indo-Japanese joint research project, we decided to digitalize BSD, to preserve original information as much as possible.

The present paper is a preliminary report on BSD digitalization. Though inputting the data has been already completed, we have found lots of work to be done. Proofreading, one of the major tasks, needs some more time. We will show what should be done besides proofreading, in order to use the data for phonemic analysis, or further morphemic studies of Santali.

2 Structure of the dictionary

BSD was reprinted by Gyan publishing house and still available, though due to the lack of preciseness of photocopying, many diacritics, which are crucial to phonetic and phonemic studies, are not always indiscernible. Fortunately the original version is available in the library of the Tokyo University of Foreign Studies, we use it for our study.

BSD consists of five volumes. The total page number is 3,406. In BSD, Santali headwords are transcribed into Roman alphabets with diacritics and superscripts, arranged in the following order, where 'h' indicates an aspirated stop.

A, Ạ, B, Bh, C, Ch, D, Dh, Ḍ, Ḍh, E, Ē, G, Gh, H, I, J, Jh, K, Kh, L, M, N, Nh, Ń, O, Ọ, P, Ph, R, S, T, Th, Ṭ, Ṭh, U, V, W, Y

The structure of each entry in BSD is as follows.

A Santali headword (a compound word, or a phrase) is followed by a comma, abbreviation(s) of parts of speech, and a period. Its English counterpart is given as the second sentence. In order to distinguish romanized Santali and English, the former is written in italic characters, the latter, upright ones. Santali sentence example(s), if any, follows the above, then followed by its English translation. Additional information such as etymological, or anything to be referred, is given within round brackets in the last of the description.

An example of headword and description is given below.

hako, n. A fish. v. a., v. m. d. Catch, get fish; v. m. Become full of fish. *H. sáko calaoena*, they went to catch fish; *h. il*, fish fin; *h. ko bãḛ̄ṛsiket́koa*, they caught fish by angling; *h. ket́koako*, they caught fish; *h. anako teheń*, they got fish to-day; *noa gaḍareko h.ḱ kana*, they get fish in this river; *noa bandre arhō̤ko h.yena*, fish have again come into this tank; *h. nãḛ̄hĩ ḍaṇḍa cele hō̤ bale ńamlet́koa*, we did not get any, neither fish nor anything. (Sakei, Besisi, Semang, Bahnar, Sue, Annam *ka*; Khasi *kha*; Nicobar *kāe*; *Muṇḍari, Birhoṛ*, Ho *haku*.)

3 Digitalization of BSD

To digitalize BSD, we need a set of transliteration rules for special diacritical symbols.[1] Following the rules, the above entry of BSD is represented as follows:

< hako >, n. A fish. v. a., v. m. d. Catch, get fish; v. m. Become full of fish. < H_0. sapˆ ko calaoena >, they went to catch fish; [... Usage Examples and Translation...]

(Sakei, Besisi, Semang, Bahnar, Sue, Annam ka; Khasi < kha >; Nicobar < ka=e >; < M_0uNDari, Birho+R >, Ho < haku >.)

As in the above, words or phrases of Santali, or of other languages which need diacritical marks are transliterated using numbers and symbols available in a keyboard, and put within angle brackets. Other parts, which are not in brackets, are in English.

In total, 39,950 entries (words, compound words or phrases) are in BSD. The total file size is 10,678,272 bytes (10MB). The whole data has already been input and the second proofreading has been almost completed. Part of the data is now available via our web site, and the rest will be so in the near future.

By providing the BSD via internet, information about Santali traditional culture and society would be accessible across the world. Our data will be sufficient for this purpose. If we are to use the data, however, in the areas of phonemics and morphology, proofreading should be done again and again, as precise reproduction of BSD spelling would be needed.

Expected cycle of our research is as follows.

Step 1: Processing the data to extract syllabic patterns found in the dictionary.

Step 2: If irregularities are found in the data, then go back to the original text to see whether these are really exceptional cases or just mistyped.

Step 3: Re-process the data to check again.

This cycle of steps will be repeated, which is an enormous amount of work, but is necessary to find regularities in the Santali phonemic patterns.

4 Results of the Analysis

Provisionary agenda for Santali phonology are:

Q1. How and why are there phonetic difference, six and eight vowel systems, among Santali dialects?

Q.2 Historically, eight vowel system of the northern dialect derived from five vowel system, or vice versa?

Q.3 How are eight vowels phonetically or phonemically conditioned?

[1] See Appendix for the transliteration rules.

These questions, related to each other, are difficult to answer. Now BSD data is available, we can consider the third question based on the large amount of data. As an initial stage of our linguistic study, headwords are extracted and classified to examine phonetic and phonemic condition of each vowels. Taking the BSD headwords as an example, we will see what we should do now.

4.1 Syllable Patterns in BSD Headwords

By representing a consonant as 'C', we extract 3446 syllable patterns and count frequencies of each pattern in BSD headwords. For example, 'CaC', that is monosyllablic /a/ flanked with whatever consonants, appears most frequently (2542 times).

Table 1 shows these order in frequency, range of frequency, number of patterns of the syllable patterns.

Order in Fr.	Range of Fr.	No. of Pat.	Order in Fr.	Freq.	No. of Pat.
No.1–8	1000 more	8	834–988	4	155
9–24	999–500	16	989–1222	3	234
25–56	499–200	32	1223–1771	2	549
57–97	199–100	41	1772–3445	1	1674
98–171	99–50	74			
172–318	49–20	147			
319–520	19–10	202			
521–833	9–5	313			
(No.1–833)	(Subtotal)	833	(No.834–3445)	(Subtotal)	2612

Table 1. Frequencies of Syllable Pattern

Hereafter, BSD's contrast [e] vs. [e̲] is shown as 'e' vs. 'e2', [o] vs [o̲], 'o' vs 'o2', respectively. The most frequently appeared patterns, Nos. 1 to 8, appear more than 1,000 times. Frequency of each pattern is hereafter given within round brackets. These are: CaC (2542), CaCaC (2244), Co2Co2C (1686), CaCa (1568), CuCuC (1542), CaCCa (1207), CuC (1195), Co2C (1173).

The above result shows that Santali prefers monosyllabic and disyllabic patterns, and that in the disyllabic cases, the same vowel tends to be repeated.

Though the whole data should be proofread carefully, patterns with low frequency should be treated most carefully. For example, the rarely appearing patterns in the right columns in Table 1 might be either loanwords from foreign languages, or simply mistyped.

4.2 'e' and 'e2' contrast

In this paper we examine the phonemic conditions for 'e' and 'e2' only. The following method of data processing could also apply for other phonemes. Among BSD entries, first we extract the patterns with either 'e' or 'e2' is included. The number of the patterns is 1,258. These patterns appears in total 12,175 times. Patterns which appear most frequently are as follows.

CaCCe (780), Ce2Ce2C (752), Ce2C (546), Ce2Ce2 (514), CaCe (401), CeCa (330), CeCCa (306), Ce2CCe2C (269), CeCaC (239), Co2Ce2 (218), CeC (205), CeCeC (193),...

It should be noted that 'e2' such as in 'Ce2C' and 'Ce2Ce2C' appears more frequently than 'CeC' and 'CeCeC' does. Minimal pairs should be checked by contrasting the patterns such as 'CeC' and 'Ce2C'.

Among the above set of entries, we extract patterns with more than two 'e's or 'e2's. There are 248 such syllable patterns. These patterns appear in total 2,946 times. Among them, the following are the most frequent patterns.

Ce2Ce2C (752), Ce2Ce2 (514), Ce2CCe2C (269), CeCeC (193), Ce2CCe2 (144), e2Ce2C (144), Ce23Ce23 (95), CeCe (78), e2Ce2 (76), e2CCe2 (34), Ce23Ce23C (29), CaeCe (20), CeCCeC (20), eCeC (20), Ce2Ce2Ce2C (18), e2CCe2C (18), Ce2CCCe2C (17), Ce2CCe2CaC (15), CeCaCe (14), CeCCe (14), Ce2CCe (11), Ce2Ce2CCe2C (11), CeCea (11), Ce2CCCe2 (10),...

Among the above subset, further extracting those with at least one 'e2', then, you get 2394 words. Again, extract from the above those with 'e', then you get those patterns in which concurrence of 'e' and 'e2' are found.

There are 74 patterns which include 'e' and 'e2' concurrence patterns. As was expected, the number of occurrence is very small; in total 149.

Ce2CCe (11), Co2Ce2Ce (8), e2CCe2Ce (8), CaCe2Ce (7), Ce2CCe2Ce (6), Co2CCe2Ce (6), o2CCe2Ce (6), Ce2Ce (5), Ce2Ce2Ce (4), Ce2CeC (4).

It should be first noticed that in case 'e' and 'e2' co-occur in a pattern, their order is always 'e2' and 'e', not vice versa.

Considering the fact that in the disyllabic cases, the same vowel tends to be repeated in Santali, the above set of patterns are exceptional, therefore we should examine carefully. Though this examination has not been done yet, a few examples of these patterns are given below.

Ce2CCe appears most frequently among the above result: 11 times.

By searching the original data, we find mẹnte 9 times, hẹnte once, jhẹdge once, respectively. Though these are listed as part of headwords in BSD, from a morphological point of view, they might further be separated as {mẹn} or {hẹn} + {te}, or {jhẹd} + {ge}. If so, these are then not exceptional disyllabic patterns, but a sequence of two separate monosyllabic morphemes.

C̲o2Ce2Ce appears in the second most frequency; 8 times.

These are: họtẹre, họtẹte, kọlẹte, mọṭhẹre, nọtẹre, nọtẹte, nhọtẹte, pọrẹte, each of which appears only once in the headword.

Again, it seems plausible to separate final syllable te and re as monosyllabic morphemes.

5 Tentative Conclusion

Now the digitalization of BSD, though further proofreading necessary. The data can be used not only for finding English meaning of Santali headwords, but for reference to Santali traditional culture and society.

In this paper, we use the data for phonemic analysis. Taking an example of the 'e' and 'e2' contrast, we analyzed the phonemic condition found in their syllabic patterns.

Though more examination is needed, it is now obvious that the 'e' and 'e2' contrast is not full-fledged vowel contrast.

1. In general, 'e2' occurs more frequently than 'e'.
2. In disyllabic patterns, both 'e2' and 'e' tend to be repeated.
3. In case 'e2' concurs with 'e', the order is 'e2' and 'e', but not vice versa.
4. In case of the above concurrence, the patterns might be analyzed further into different morphemes.

Bibliography

Anderson,GDS (2006): Santali, pp.749-751. *Encyclopedia of Languages and Linguistics*, Elsevier.

Bodding, P. O (1932-1936): *A Santal Dictionary*, 5 vols. Oslo: Norwegian Academy of Science and Letters.

Minegishi, Makoto and Murmu, Ganesh (2001): *Santali Basic Lexicon with Grammatical Notes,* 246pp. ILCAA, Tokyo University of Foreign Studies, Tokyo.

This research has been supported in part by the project "Grammatological Informatics based on Corpora of Asian Scripts" (2001–2005), in part by the project "Corpus-based Linguistics and Language Education" (Global COE Program: 2007–), granted by MEXT, and in part by the project "Multilingual Concierge – An Interface for the Next Generation" granted under the "Strategic Information and Communications R&D Promotion Program" (SCOPE) of the Japanese Ministry of Internal Affairs and Communication.