1 Introduction

1.1. Preview

The purpose of this chapter is to provide a general description of the work that has been chosen for the present research and related issues. This work is concerned with the description of grammatical categories of Madhav Kandali's Rāmāyana, an old Assamese text attributed to the 14th century AD (cf. §1.2). This was period of great significance in the history of the Assamese language, during which the first phase of renaissance in the fields of religion, culture and language was taking place. The state of the Assamese language is clearly reflected in the works of Madhav Kandali and a few other court poets of this period. The phrase in italics is intended in two senses: First, these works suggest that the Assamese language existed before the period when Rāmāyana was written. Moreover, it also appears that the language of that period had not taken a distinctive form as yet. It is believed that the spoken form of the Assamese language began to develop and spread among the people during the period between the fifth and seventh centuries AD. The evidence for this comes from the copper plate grants and rock inscriptions found in the fifth century AD, beginning with the Umācal rock inscription of Surendravarman up to the Tezpur plates of Vallabhadeva. Though the language of these copper plates and inscriptions is Sanskrit, certain linguistic irregularities and errors seen in those inscriptions lead linguists to trace the existence of the local Prakrit language side by side with the literary Sanskrit language of the royal court (Sharma 2007: 264), and the influence of this Prakrit on the Sanskrit language of those epigraphs. In this regard, Barua observes:

The archaic forms crept in these records and held their rightful place through the inadvertence on the part of the local composers, or that of the local scribe-engravers or that of both, in spite of the conscious attempt made for producing the legal documents, the land grants, in authentic Sanskrit. (1947; cited in Sharma 2007: 265)

The existence of local Prakrit at that period is supported by the comment of the Chinese pilgrim Hiuen Tsang when he visited Kāmrupa in 643 AD. According to him,
“their speech differed a little from that of Mid-India” (Watters Vol. II. P. 186). By “their speech”, he undoubtedly referred to the people of Kāmrupa, because at the time of he visited Kāmrupa, the western boundary of Kāmrupa was found to extend to the river Karatoya in North Bengal and according to the description of Watters (Vol. II. P. 186, 187) “the pilgrim crossed a large river and came to Ka-mo-lu-p’o” (Kāmrupa). It is assumed by him that the large river may be the river Karatoya of the western boundary of Kāmrupa kingdom (Cited in Kakati 1941: 5). Sharma (2007: 265) assumes that the Assamese language emerged from this local Prakrit, and further developed in the language of Caryās, written on palm leaves (tāl pāṭ) by Buddhist Siddhāchāryas hailing from the different parts of ancient Kāmrupa. The language of Caryā is not completely Assamiya, and is surmised to be "late apabhramsa, and represents the formative period of the NIA languages including Assamiya" (Goswami and Tamuli 2003: 397). According to Sharma (2007: 266), the language of Caryās contains definite traces of the Assamese language in its formative stages, whereas the traces of local Prakrit found in the copper plates and inscriptions mentioned above contain only remote and indirect indices of Assamese. Apart from Caryās, the traces of the Assamese language in the formative period can be seen in the evidence of inscriptions found at Ambari, Guwahati, which “contains the specimen of the Assamese language prevalent in the middle of the thirteenth century” (Sharma 2007: 268). Another specimen of the formative period of Assamese is found in the language of Kṛṣṇa Kirtan. Regarding the language of this text, Kakati observes,

The Kṛṣṇa Kirtan presents a mixture of dialects which have not yet started on courses of independent development but the beginnings of which seem to have been well laid down. If it be allowed to represent the Pre-Bengali and Pre-Assamese dialect groups by the unknown quantity X, then it may be said that Kṛṣṇa Kirtan preserves specimens of latter-day X-dialects which in later times developed into distinctive Bengali and Assamese languages. (Kakati 1941: 12)

It was only in the latter half of the 13th century AD or the first part of the 14th century AD that the Assamese language began to take a mature form, which is reflected in the Rāmāyana of Mādhav Kandali.

The present research is intended to serve five purposes. Firstly, it seeks to contribute to our understanding of the nature of development of the Assamese language: it is hoped that the results of such investigation can usefully contribute to an overall
impression of how the Assamese language has arrived at its present state of evolution. Second, this research will facilitate the comparison of the language of the 14th century, with varieties of Assamese of other time periods with regard to the exponency of grammatical categories. Third, such a study can be easily extended in the direction of (i) comparisons with cognates such as Bengali on the one hand, and (ii) comparisons with the language of other writers considered as belonging to the same period. Fourth, it is designed to serve as a foundation for further study of language materials from the perspective of morpho-syntax, since most of the studies carried out in Assamese so far, are basically morphologically based (cf. §1.5). Finally, it is designed to serve as a foundation for corpus-based studies in Assamese using computer-readable corpus data and for creating a set of tagging guidelines and tag set for analyzing linguistic data of old Assamese for future researchers.

This chapter is divided into eight main sections including the present one. §1.2 provides a brief description of Mādhav Kandali, the author of the text, and his times. The specific aspects of Rāmāyaṇa selected for analysis are briefly indicated in Section §1.3. A review of relevant work by other scholars is undertaken in Section §1.4. §1.5 focuses on the methodology of research, with Section 1.6 presenting a discussion of the theoretical framework in terms of which this study has been carried out. Certain issues that are encountered while analyzing the data are examined in §1.8. §1.7, the final section presents the overall structure of the present work.

1.2. Mādhav Kandali and his times

This section provides a brief account of contemporary poets with Mādhav Kandali and their contribution to Assamese language and literature. The kings who patronized their literary endeavours are also considered as a means of determining the time of Mādhav Kandali from the reference of those kings on the one hand and to determine the time frame of the language of MKR on the other. In addition, such a background is also helpful in finding out what the status of the Assamese language was during that time and which regional variety of the language was adopted to compose this text. Such information is potentially useful in accounting for the behaviour of some forms and functions of this text. More generally, this discussion is intended to provide a general
idea of the state of the Assamese language and literature during the time of Mādhav Kandali.

As stated above, the earliest Assamese literature developed in the hand of a few court poets including Mādhav Kandali. Hem Saraswati was the first poet of that time who wrote *Prahlād Charitra*, a small poem, at the behest of King Durlabhnhārāyana of Kamatā. Thus, from that point of view, Prahlād Charitra is considered as the earliest Assamese writing where a distinctive form of the language is found to exist. Haribar Bipra, another court poet who mentions Durlabhnhārāyana as his patron, composed the *Aswamedha Parva* of the *Mahābhārata*. However, Kakati does not want to consider Haribar Bipra as contemporary of Durlabhnhārāyana, but as the contemporary of Indranārāyana, the son of Durlabhnhārāyana (Kakati 1941: 13). Indranārāyana was the patron of another poet named Kaviratna Saraswati, who wrote *Jayadratha Vadha*.

Mādhav Kandali, the composer of Rāmāyana, on which the present research is based, flourished at a period later than that of the poets mentioned above. At the behest of King Mahāmānīkya, Kandali rendered the entire Rāmāyana into Assamese verse form from the original Sanskrit *Rāmāyana* written by Bālmiki. Kakati refers to Kandali as a most considerable poet (Kakati 1941: 14) of this period, because his Rāmāyana is a remarkable contribution to the Assamese language and literature. Rudra Kandali, another court poet was a contemporary of Mādhav Kandali. He translated *Drona Parva* following the royal wish of King Tāmradhvajā.

The writers of this period did not clearly indicate their place of birth and time anywhere. From the mention of their names in their writings along with the names of the kings under whose royal instructions they undertook the literary works cited above, the time of their existence and their place of birth were sought to be inferred by such scholars as Kakati (1941: 14); Barua (1933: 213); Medhi (1936: 64, 70); Barua (1951: 156); Goswami (*Bānhi* Vol. XVIII No. 5; cited in Barua (1933: 213, 214), Gait (2008). Based on the time and the geographical and political boundary of the kingdoms, the scholars tried to figure out the approximate time and place of the poets of this period.

It is known from the history of Kāmrupa, that after the end of the Pāla dynasty in Kāmrupa in 12th century AD, the Kachāris became powerful and spread their kingdom towards the western boundary of the present district of Kāmrupa. Being unable to stand up to the powerful Kachāri kings, the kings of Kāmrupa were forced to demarcate the
boundary of their kingdom by shifting the capital of Kamrupa from Gauhati further west to Kamatapur and announced themselves as kings of Kamatã. Instead of the large geographical boundary of the old Kamrupa, the Kamatã kingdom included only the Goalpara and the Kamrupa districts of the modern Assam Valley along with some portion of North Bengal and Mymensing to the east of the Brahmaputra (Barua: 1933: 160). Durlabhnârâyana, the patron of Hem Saraswati who is credited with the earliest writing in Assamese, was reigning at Kamatapur around the latter part of the thirteenth century across the boundary of Kamatã between the Bar Nadi and Karatoyã (Medhi 1936: 63). However, scholars do not seem to agree with each other regarding the time and place of birth of Mahämânîkiya, the patron of Mâdhav Kandali who composed the Râmâyana. The poet himself mentions in Lankâkãnda that he composed the Râmâyana at the behest of Mahämânîkiya, Barãha Râjâ: 

\[
\begin{align*}
\text{kabiräjä kandali ye,} & \quad \text{ämäkese buliwa}ya, \\
\text{karilohon sarbbajana bodhe.} & \\
\text{rämäyäna supayära,} & \quad \text{śrimahä mäpîke ye,} \\
\text{barãha räjära anurodhe.} & \\
\text{säta kända rämäyäna,} & \quad \text{pada bandhe nibandhilo,} \\
\text{lambhä parihari särodhrta.} & \\
\text{mahämänîkara bole,} & \quad \text{käbyaras kicho dilori,} \\
\text{dugdhaka mathile yena ghṛta} &
\end{align*}
\]

(Lañ, Ch. 56, V. 24 & 25)

'We are known as Kabiräjä Kandali. Following the order of Mahämânîkiya, the king of Barãha, we composed the seven kändas of Râmâyana in brief, avoiding lengthy descriptions, by adding some literary flavours, in a simple style, as in the process of producing butter by churning milk, so that the general people understand the story of Râmâyana very well.'

Unfortunately, Kandali does not mention anywhere the location of the place where Mahämânîkiya reigned and to which community he belonged.

In fact, apart from mentioning his name in the middle of the text, the poet does not give out any clue about his identification. He mentions his name eighteen times in the entire text, and these references are the only source that indicates that he was the writer
of Rāmāyana. Two out of the numerous verses (slokas) that occur at different places of the whole Rāmāyana, are cited below as an evidence of his self-identification.

\[ \text{kari śīrogata rāma dewara caraṇa.} \]
\[ \text{mādhava kandali biracilā rāmāyaṇa.} \]

‘By worshipping to the feet of Lord Rama, Mādhav Kandali composed the Rāmāyana.’

(Ar, Ch. 1, V. 6)

\[ \text{mādhava kandali bipre, tāhāna caraṇa smari,} \]
\[ \text{karilanta ślokaka uddhāra.} \]
\[ \text{rāmara caraṇa bīnā, ānagati nāhi herā,} \]
\[ \text{jānibāhā mane dṛśha kari.} \]

‘By remembering his (Rama) feet, the Brahmin Mādhav Kandali composed the slokas of the Rāmāyana. Be sure that there is no other way of living other than worshipping Rama.’

(Su, Ch. 41, V. 9)

The second verse carries another information about Mādhav Kandali, namely, that he was a Brahmin scholar. The name Kabirāja Kandali mentioned above suggests that he may have been the greatest among poets and hence, this title may have been given to him by his patron in recognition of his scholarly knowledge.

In addition to such self-identification, the comment of Sankardeva, who was born in 1449 AD and who refers to Mādhav Kandali as an eminent predecessor, provides an important clue regarding Mādhav Kandali. He makes the following remarks in the Uttarākānda chapter of MKR that is attributed to him:

\[ \text{pūrba kabi apramādi, mādhava kandali ādi,} \]
\[ \text{biracilā padē rāma kathā.} \]
\[ \text{hastīra dēkhiyā lāda, saśā yēna phārmārga,} \]
\[ \text{mōra bhalā tēhṇāya awasthā} \]

‘Unerring predecessor poets such as Mādhav Kandali and others composed the story of Rama in verse. My state is that of a hare who
breaks wind at the sight of an elephants’ turds, as if to challenge the elephant.'

(Utt, Ch. 19, V. 6)

Further evidence is found in the biographical work on the Assamese Vaisnava gurus (Lekharu 1952: 26; cited in Saikia (Bora) 1985: 24). According to this biographical work, Raghavacarya, a student of Mādhav Kandali came to inspect the tol1 of Mahendra Kandali who was the teacher of Sankardeva and this reference helps to locate the time of Mādhav Kandali as being prior to that of Sankardeva.

On the basis of such information, scholars have attempted to fix the place and time of Mādhav Kandali in the following manner.

According to Medhi (1936: 70), Mādhav Kandali composed the Rāmāyana under the patronage of King Māhā Mānikiya of Tripurā around the middle of the fourteenth century AD. He describes the political boundary of ancient kingdom of Tripurā (Tippera) including the Kapili Valley, along with North Cachar Hills and Cachar and continues as follows:

About the fifth century A.D. Drikpati, a scion of the Tripurā Rāj family, who succeeded to the throne of his father-in-law, the King of Heḍamba, invaded the Kapili valley and drove the King of Tripurā to Carhar. About the thirteenth century A.D. Ratnapha, King of Tripurā, first assumed the surname of “Mānikiya”. Since then all succeeding Kings of Tripurā and also probably the Kings of Jayantia and Heḍamba assumed that surname. Ratnapha’s great-grand-son was “Māhā-Mānikiya”. (Medhi 1936: 64)

According to Medhi, Mādhav Kandali translated the Rāmāyana following the instructions of this Māhā Mānikiya of Tripurā.

Barua (1933: 213) places Mahā Mānikiya in the middle of the fourteenth century AD. While talking about the kingdom of Mahā Mānikiya, Barua refers to the Kapili Valley. From their comments, it is apparent that both agree in locating Mahā Mānikiya in the Kapili Valley or somewhere near. But like Medhi, Barua does not introduce Mahā Mānikiya as a king of Tripurā, but as a king of the Barāhi Kachāris. Goswami too states that Mahā Mānikiya was a king of the Barāhi Kachāris who reigned around the middle of the fourteenth century at Dimāpur (cited in Barua 1933: 213, 214). He also observes

---

1 Like a present day school, Tols were established by different Gurus (teacher) for teaching Sanskrit to students in ancient time. In some place, this kind of Sanskrit tol exists in the present time also.
that “Dersong Pha, great-great-grandson of Maha Manikya was the contemporary of the Ahom king Suhunnung alias Dehingia Raja” (1933). According to Bardalai (1899), the first editor of the printed version of Madhav Kandali’s Rāmāyan, Mahā Mānikiya was a Kachāri king of Heđamba who reigned about the middle of the fourteenth century AD. Before commenting any further on these two points, it would be relevant to dwell a little on the political scenario of Kāmrupa around the 13th century.

The old Kāmrupa kingdom was divided into different kingdoms of various sizes after the end of the Pāla dynasty. During the first part of the thirteenth century, the Ahoms entered into the eastern corner of the Brahmaputra Valley and started to reign there. In the east of the Subansiri and the Disang (Gait 2008: 39), the country seems to have been ruled by a line of Chutiya kings. According to Barua (1933: 161), the Chutiya kings seem to have reigned on the north bank of the Brahmaputra. The south bank of the Brahmaputra seems to have been ruled by the Kachāri kings. To the west of the Kachāris and the Chutiyaas, the country seems to have been ruled by a number of Bhuiyas.

From the descriptions of Gait (2008: 300, 301), it is known that the political boundary of the Kachāri kingdom included the south bank of Brahmaputra from the Dikhu (present Dikhou) to the Kallang along with the valley of Dhansiri and the present state of North Cachar sub-division around the 13th century AD. In the account of the Ahom invasion against Kachāri kings, Gait mentions Dimapur as a capital of the Kachāris. He observes,

The Ahoms followed up their victory and, ascending the Dhansiri, penetrated as far as the Kachāri capital at Dimapur on the Dhansiri, forty-five miles south of Golaghat.

(2008: 301)

Gait further states, “The ruins of Dimapur, which are still in existence, show that, at the period, the Kachāris had attained a state of civilization considerably in advance of that of the Ahoms” (2008: 301). Gait’s account seems to support Goswami’s observation stated above.

The name Heđamba, mentioned by Medhi (1936: 64) and Bardalai (1899) may be connected to the Hidimba mentioned by Gait (2008). In this regard, Gait remarks,

The Kachāri king at that time was styled “Lord of Hidimbā”. After this
time, the name Hidimbā or Hiramba frequently occurs in inscriptions and other records, but there is no evidence of its use by the Kachāris at any earlier period. It has been suggested that it had long been the name of the Kachāri Kingdom, and the Dimāpur is in reality a corruption of Hidimbapur, but it seems more likely that Hidimba was an old name of Cachar, which the Brahmins afterwards connected with the Kachāri dynasty. (P. 303)

From this evidence it appears that whatever may have been the source of Hidimbā or Heḍamba, all the scholars seem to agree that this was a name of Kachāri kingdom.

Kakati (1941: 14), on the other hand, mentions that Mādhav Kandali rendered the whole of the Rāmāyana at the behest of King Mahā Māṇikya, a Kachāri king of Jayantāpura. However, he also does not want to place Mādhav Kandali later than the fourteenth century AD. However, Barua (2003: 156), places Mahā Māṇikya in the thirteenth century.

From the above comments, it can be seen that there is hardly any difference of opinion about the time of Mādhav Kandali. Except for Barua (2003), all other scholars agree to place him in 14th century AD.

On the other hand, scholars do not seem to agree on Kandali’s place of birth. However, from the range of opinions expressed by these scholars it can be assumed that Mahā Māṇikya was a king of the Kachāri kingdom within the political boundary described above, not a king of Tripurā as mentioned by Medhi. Gait’s comment in this connection is relevant:

Up to A.D. 1603 ...... the Kachāri kings held the greater part of the Nowgong district and the North Cachar Hills and gradually extended their rule into the plains of Cachar............there is a tradition that it was formerly included in the Tippera kingdom, and was presented by a king of that country to a Kachāri Raja who had married his daughter, about three hundred years ago.

(2008: 304)

It is likely that Gait’s comment had motivated Medhi to include Kapili Valley along with North Cachar Hills and Cachar within the boundary of ancient Tripurā kingdom, and to consider Mahā Māṇikya as a king of Tripurā. Moreover, the surname Māṇikya used by the Tripurā may have led Medhi to this conclusion. However, since the surname Māṇikya was also used by the kings of Jayantāpura and Heḍamba, it is plausible that Mahā Māṇikya was introduced as a king of the Kachāri kingdom. Moreover, it seems far fetched to assume that a reputed Assamese scholar such as Mādhav Kandali is likely
to have composed an Assamese Rāmāyana under the patronage of Tripurā king despite the existence of as many as five kingdoms within the boundary of Old Kāmrupa as evidenced above.

The above accounts are all based on the political scenario of Kāmrupa during 13th and 14th century AD. However, the place of birth of Mādhaṅ Kandali does not appear to have been considered from the perspective of linguistic evidence. There is no doubt that linguistic evidence often helps to identify areal information. As far as the place and time of Mādhaṅ Kandali is concerned, the language of the Rāmāyana appears to be closer to the language of present Kāmrupa including Nalbari, Goalpara and Bongaigaon as well as to the language of North Bengal (Koch Bihār) where the old Kamatā kingdom existed, than to the language of the places in which the Kachāri kings ruled. Some supporting linguistic evidence is given below.

The similarity between the language of Rāmāyana and the language of the western part of present Kāmrupa as well as the languages of Koch Bihār is mainly lexical. However, to some extent the correspondences are found at the morphological and syntactic levels as well.

(i) Lexical similarity

kena ‘how’  keman ‘how’  kichu ‘nothing’
koṭhā ‘where’  kāche ‘near’  ethāne ‘here’
ethāna ‘here’  ehna ‘he. PRX’  ite ‘now’
etadura ‘this far’  ihāra ‘he. PRX’  tāhāra ‘he. DST’

(ii) Morphological similarity

-khāna ‘CLF’ (dhānukhāna) -ka ‘DIRT’ (mithilā-ka ‘to mithila’)
-tda kari ‘than’ (mota kari ‘than me’)

(iii) Syntactic level

mithilā-ka lāgi ‘to Mithila’  mo-ra janya ‘for me’
Mithila-DAT to I-GEN for
kar-ibāka lāgi ‘to do’
khar NF to

From this evidence, Mādhaṅ Kandali can be located somewhere in the old Kamatā Kingdom. However, such evidence is hardly conclusive. Moreover, it is not an area that falls within the scope of investigation of the present researcher. Hence the present
discussion has been limited to a consideration of the observations of other scholars on the matter of the place and time of Mādhav Kandali. Kandali's contribution to Assamese literature is taken up next as this will serve to highlight some issues relating to MKR.

1.2.1. Kandali's contribution to Assamese Literature

Mādhav Kandali was the first composer of Rāmāyaṇa in Assamese. In fact, he was the first poet of North-East India to have written Rāmāyaṇa in the Desi Bhasa. He translated the entire Sanskrit Rāmāyaṇa of Bālmiki into Assamese. Kritivāsa's Bangla Rāmāyaṇa and Tulsi Dasa's Hindi Rāmāyaṇa were written many years after Mādhav Kandali composed the Assamese Rāmāyaṇa. However, the Tamil Rāmāyaṇa of Kambana and the Rāmāyaṇa in Telegu and Malayalam are much older than the Rāmāyaṇas composed in North India. Among the three recensions (Pātha) of Bālmiki Rāmāyaṇa found in different regions of India namely Gauḍiya or prācyā, Udicya or paschimottar (North-western) and Daksinatya (Southern) (Sarma 2007: 28), Mādhav Kandali adopted the source of Rāmāyaṇa from the Gauḍiya Pāth (Saikia (Bora) 1993: 1).

Mādhav Kandali composed the seven kāndas of Rāmāyaṇa - Adikānda, Ayodhyākānda, Aranyakānda, Kiskindhākānda, Sundarakānda, Laṅkākānda, and Uttarakānda. He makes a statement to this effect at the end of Laṅkākānda:

\[
\begin{align*}
\text{kabirāja kandali ye,} & \quad \text{āmākēse buliwaya,} \\
\text{karlōhōṁ sarbajana bōdhē.} & \\
\text{rāmāyaṇa supayāra,} & \quad \text{śrimāhā māṇikē ye,} \\
\text{barāha rājāra anurōdhē.} & \\
\end{align*}
\]

'Ve are known as Kabirāja Kandali. We composed the Rāmāyaṇa following the order of Mahā Māṇikya, the Barāhi King.'

(Laṅ, Ch. 56, V. 24)

However, it is believed that the first and the last kāndas of his Rāmāyaṇa do not exist now. It is further believed that these two kāndas got lost during the battle of the Kachāris and the Ahoms (Bardalai (1899, cited in Sarma 1998: 11). The Adikānda and Uttarakānda found today are composed by Mādhavdeva and Sankardeva at a later
period. This is known from a reference of self-deprecation of Sankardeva when he started to write the lost version of Uttarāṅkāṇḍa in order to lend completeness to the seven kāndas of the Rāmāyana written by his former unerring poet Mādhav Kandali:

\[
\begin{align*}
\text{karayōrē buddhalōka,} & \quad \text{bōlō kṣēmiyōka mōra,} \\
\text{āmāra ayōgya ītō karmma.} & \\
\text{padata dūsaṇa pāi,} & \quad \text{nindībēka nuyuwāi,} \\
\text{mahanta jānara kṣāmā dīarmacīna.} & \\
\text{pūrbba kabi apramāḍt,} & \quad \text{mādhawa kandali ādi,} \\
\text{biracilā padē rāma kathā.} & \\
\text{hastīra dēkhiyā lāđa,} & \quad \text{śasā yēna phārēmārga,} \\
\text{mōra bhaila tēhpāya awasthā.} & \\
\end{align*}
\]

'O wise people, with folded hands we request you to forgive us for composing the Rāmāyana — a task for which we are not exactly eligible. Unerring predecessor poets such as Mādhav Kandali and others composed the story of Rama in verse. My state is that of a hare who breaks wind at the sight of an elephant.'

(Utt, Ch. 19, V. 5, 6)

Some scholars claim that apart from Rāmāyana, Mādhav Kandali wrote another puthi named Devajit (Goswami; cited in Sarma 1998: 30); (Sarma 1996: 69).

The foregoing remarks were intended to highlight the important place that the Rāmāyana occupies in Assamese language and literature. This was sufficient to motivate the present researcher to undertake research on this text. The next section is concerned with an exposition of the type of linguistic research that has been conducted on this text.

### 1.3. Topic of research

The present research is concerned with grammatical categories found in Mādhav Kandali’s Rāmāyana (MKR). Although this study will be largely synchronic in nature, reference will be made to historical processes wherever relevant and helpful. Such an
approach will often be found to offer a more meaningful account of the synchronic distribution of forms being discussed.

It may be noted that there are two ways in which grammatical categories have been defined in the linguistic tradition. One way considers the major word-classes (cf. §2) such as nouns, verbs, adjectives, and adverbs as grammatical categories. A second approach considers the features associated with particular word-classes, such as case and number in nouns and tense and aspect in verbs, as grammatical categories. The second approach is normally adopted by linguists, and this has also been adopted for the purpose of the present research.

Accordingly, grammatical category is defined as a linguistic category which is associated with major word classes for the purpose of carrying different grammatical information in order to serve human communication successfully. Its presence in all languages of the world is as important as the presence of lexical categories such as nouns, verbs, adjectives etc. It is true that all languages may not share the same grammatical category or the distribution of a particular grammatical category may not be similar across the languages of the world, but all languages must have certain grammatical categories to carry grammatical information. For example, the noun of English exhibits only one grammatical category, that of number, but the nouns of Assamese are affected by three grammatical categories: number, case and classifiers. Similarly, in English, the grammatical category of comparison is represented on adjectives in two ways: by using comparative and superlative suffixes, and by using a set of closed class words. As against this, in Assamese it is expressed by using suffixes alone. The verbs of English and Assamese both exhibit the categories of tense, aspect, mood, and person. However, person does not play much role in English syntax whereas this category plays an important role in Assamese syntax. Moreover, Assamese has an additional set of category called personal deixis. Similarly, the category of visibility is a unique feature for North American languages (Trask 1999: 112) which is rare in other languages. In short, it can be said that grammatical category plays an important role in the organization of grammar in any language. The presence of a grammatical category accounts for why members of a particular word class have different forms. MKR has a rich array of word forms for which the perspective of grammatical category is a promising one and this is the reason for choosing it as a topic of the present research.
Based on their role in the organization of grammar, grammatical categories are divided into three types: Inherent, Relational, and Agreement. These will be discussed in detail in the next chapter after we have covered word classes – an area with which the notion of grammatical category is closely related.

Though almost all major word classes take grammatical categories, the scope of the present research has been delimited to the word-classes of verbs and nouns as they are more basic elements of sentence structure. More importantly, these two word classes play a major role for the grammatical categories found in MKR as will be evident from their treatment in different chapters of this work.

While we shall not enter into any further discussion on grammatical categories in this section, two other aspects relating to the language of MKR will be taken up for discussion below – these are the study of manuscripts and the phonology of MKR.

1.3.1. The study of manuscripts of MKR

Though Mādhav Kandali’s Rāmāyana is available in the form of manuscripts, it was decided not to use them as the basis for the present research project. There are two reasons for this decision. The first reason relates to the purpose of the present project. As stated earlier, one of the main purposes of this project is to design a foundation for a corpus-based study on Assamese in a computer-readable form. To serve this purpose, a printed version of Mādhav Kandali’s Rāmāyana was regarded as a better source. From a methodological point of view, it can be a formidable task to convert a text from manuscripts into an electronic form. One reason for this is that manuscripts can be illegible in sections, requiring a corpus creator to reconstruct what the writer might have written. Moreover, it may contain spelling inconsistencies, colons and dots marking prosodic groupings etc. (Meyer 2002: 78). To arrive at a conclusion about the status of a word displaying different spellings, the textual criticism of all the manuscripts of a text is needed. Towards this end, it is first necessary to carefully study all the manuscript versions. Under such circumstances, if a specific text has several manuscript versions, it would take a lot of time to type in all the manuscripts variants in order to convert them into a computer-readable form.
In addition, if the corpus analyst needs to create her own corpus for undertaking any linguistic research, then the burden of work would increase twofold. This is because in her role as a corpus compiler, the analyst has to keep her attention focused on methodological concerns. If the analyst had decided to create her own corpus from manuscripts, which in the present instance is a historical corpus, she would have had to play the additional role of an editor as well. As an editor, she has to normalize the text to be computerized, taking decisions about which variant spellings to consider, which diacritics to include, and so on. Thus, it is not feasible to handle the two tasks of compiling a corpus from old manuscripts and conducting research on it at the same time. The manuscript of MKR can be cited to illustrate this situation. According to Saikia (Bora), several manuscripts of Mādhav Kandali’s Rāmāyana are scattered in various places of Assam (1985). The careful study of all these manuscripts would require a lot of time and effort. It is itself a large project in its own merit. Within the span of the five-year project of the present researcher, it has not been possible to undertake all aspects relating to Mādhav Kandali’s Rāmāyan. Furthermore, the goal of the present research is not to undertake a textual criticism of Mādhav Kandali’s Rāmāyan. Rather, as stated earlier, it is an attempt to investigate certain aspects of the grammatical structure of MKR using its data in a computer-readable form. However, while selecting the corpus for analysis, care has been taken to ensure the authenticity of the source.

Four versions of MKR have been located. Theoretically, a corpus compiler could create a corpus containing all these four versions, and then let the analyst decide which version(s) to analyze, or provide some kind of interface allowing the analyst to compare the various versions. However, due to the enormous amount of work involved in computerizing all these versions of MKR, it was much more practical to computerize only one version. For this reason, the edited version of Sharma (1988), in conjunction with the Bardalai (1899), has been chosen to convert to computer-readable form. The advantage of choosing an edited version as the basis to computerize is that such a version was already created by someone who went through the various manuscript versions, made decisions concerning how variations in these versions ought to be reconciled, and in a sense produced, a version that never existed. Of these two versions, the one edited by Bardalai is the first printed version of Rāmāyana.
1.3.2. The phonology of MKR

The phonological study of MKR has been excluded from the scope of the present project. Since MKR is an orthography-based text, spoken evidence of the language of the text is not available to us for phonological analysis. Hence, it is impossible to get a clear picture of phonological structure of that language of that time. Moreover, the manuscripts of MKR that are available in the present time cannot be claimed as original. These are likely to have been rewritten many times by different scribes, giving rise to the possibility that the shapes of particular forms might have undergone some change in the hands of those scribes. Further, the methodology of reading the manuscripts is a matter of investigation in its own merit: to what extent these manuscripts were read scientifically or whether the scribes were properly trained in this regard are important issues that have a bearing on the matter. Without investigation of these issues, it is hard to deal with the phonology of any language just on the basis of orthographical structure.

However, it is possible in principle to undertake this type of study through the reconstruction of proto-forms based on the phonological structure of modern Assamese along with the different varieties of the language that exist today, as well as other close cognates of Assamese. Without a comparative study of such present forms of modern Assamese and its close cognates, the phonological structure of the language of the time of MKR cannot be figured out.

Even though it was stated earlier that the study of the manuscripts of MKR and its phonology are both related to the language of Rāmāyana, their relevance in this case is only in terms of their potential for contributing to individual linguistic research. Neither aspect had much to say about the nature of the language of the Rāmāyan which we shall now consider in the next section (§1.4).

1.4. The language of Rāmāyana

The language of MKR belongs to the poetic genre. Making due allowances for differences in genre, it can be assumed that Kandali modelled his writing on the spoken
form of the language of his time. Two types of spoken form are evident from the pattern and structure of the use of words, and from the mode of sharing thoughts and emotions in MKR. One form appears to be used by the people from the higher stratum of the society, and the other by those belonging to a lower stratum. It is reasonable to assume that for his work, Kandali adopted the spoken variety of the place where he lived, and modified this language with suitable imagery and different metrical patterns, drawing on his knowledge of the Sanskrit language in the process. The evidence of spoken language is found all over the text. In particular, from the use of different pronominal forms, it is evident that Mādhav Kandali composed this text by using the spoken variety of present-day lower Assam including Kamrup, Nalbari, Barpeta, and Goalpara districts. Furthermore, a suffix such as -ka, syntactic expressions like -ka lagi found extensively in MKR are still used in some varieties of lower Assam (cf. §1.2). Both devices are used to indicate direction in MKR as well as in modern Assamese.

The metrical and other aesthetic requirements of the poetic genre appear to have induced the language of MKR to assume various shapes.

In the first place, even though Kandali adopted the spoken variety of Assamese for his work, it is, however, true that the selection of some words is more literary as against words from the everyday colloquial language. It is quite evident that people did not speak with each other in this style of poetic form. The use of alankār, rupak, upamā and many formal words set the text apart from the colloquial forms of the language of that period.

Secondly, there seems to have been a frequent change in the word order for the sake of adjusting the rhyming pattern in the text. As a result, the word order of the language of the 14th century cannot be figured out from the text.

Thirdly, some suffix-like elements are found to be attached to words occurring at the end of the line only for the purpose of adjusting the rhyme with the word occurring in the previous or subsequent line, as illustrated in (3) below.

2 Though the spoken evidence of this period is not available to us, from the use of some forms of Rāmāyana (with a slight modification) that are quite close to the present-day varieties of Assamese spoken in lower Assam, it can be assumed that Kandali drew upon the spoken variety of the time while translating the Rāmāyana.
In the first example, the word \textit{sare} occurring in the end of the line seems to take the \textit{-e} suffix only for adjusting the rhyme with the word \textit{lankesbare} in the subsequent line.

Similarly, the word \textit{sagaranta} occurring in the last line of the second example seems to take the element \textit{-anta} only for adjusting the rhyming pattern with the word \textit{uranta} occurring at the end of the previous line. Actually, this suffix is the locative \textit{-ta}. The element \textit{-an} seems to have been added here just for adjusting rhyming pattern. Thus, the rhythmic nature of the text seems to impinge upon vastly on the pattern/usage of suffixation. Many suffixes are found to have been used randomly for adjusting the rhyming pattern only. The use of the \textit{-e} suffix to mark intransitive and transitive subjects can be considered as an example. This suffix occurs optionally with both transitive and intransitive subjects without any underlying regularity. In such situations,
a generalization has been attempted by comparing it with its later developments, close cognates and sister languages (§7).

In addition, the language of Rāmāyana seems to have been greatly affected by orthographic corruption. Many spelling mistakes and inconsistencies can be seen in the entire text, making it difficult to obtain a clear picture of the orthographic system of Assamese of that period from this data. The reasons for such spelling mistakes and inconsistencies have been discussed in relevant places in the subsequent chapters.

Finally, some words from cognate Bangla are found in MKR. This is to be expected, since the geographical as well as the political boundary of old Kāmrupa included the northern part of Bengal, the present Koch Bihar at the time when MKR was written. Moreover, it is assumed that Mādhav Kandali composed this text at a location somewhere in present-day lower Assam. Linguistic evidence also supports this assumption (§1.2). Due to this reason, Bangla words might have entered freely into the text.

It is true that the familiarity of the author with the Sanskrit literary tradition has lent a certain classical flavour to the language of the work, but Kandali did not use the style of Sanskrit in his translation. He wrote the text in an easy and simple, yet sublime style to cater to the needs of spiritual leaders as well as for the religious enlightenment of the mostly illiterate masses.

\[
sāta kānda rāmāyaṇa, \quad \text{pada bandhe nibandhilo},
\]
\[
lambhā parihari sārodhṛta.
\]
\[
mahāmāṇikara bole, \quad \text{kābyarasa kicho dilohī},
\]
\[
dugdhaṁkara mathile yenā gṛhṛta.
\]

‘Following the order of Mahamanikya, the king of Baraha, we composed the seven kandas of Rāmāyana in brief, avoiding lengthy descriptions, by adding some literary flavours, in a simple style, as in the process of producing butter by churning milk, so that the general people can understand the story of Rāmāyana very well.’

(Lañ, Ch. 56, V. 25)

Due to some of the nature of the language, the present researcher had to face certain issues during data analysis. These will be discussed in §1.8. The next section (§1.5)
deals with earlier studies on the language of MKR relating to the present topic of research and offers a critical view on such work.

1.5. Earlier studies on the language of MKR

Only a few researchers have done their doctoral research on the language of MKR. Out of these, the works of Thakuria (1983), Saikia (Bora) (1985) and Bardalai (2005) are noteworthy. The focus of Thakuria (1983) is on a text-critical and historical study of the Rāmāyana of Mādhav Kandali. Saikia (Bora) (1985) and Bardalai (2005), on the other hand, are concerned with the linguistic aspects of MKR.

Other scholars have done their research on various aspects of the language of the 14th century, but none has specifically addressed issues relating to MKR. One of them is Borah (2000), which is concerned with the verbs occurring in the Assamese literature of the pre-Sankardeva era i.e., with the language of the 14th century. Several other works, notably (Kakati 1941), Sarma (1961), (Neog 1962), Goswami (1970), Goswami (1982), and Sarma (1996) also contain some general reference to the language of MKR. Among these, we will review only those areas of their work that are specifically related to the topic of present research, and more generally, on the language of MKR.

Saikia (Bora) (1985) is the first extensive study of the linguistic aspect of MKR. The thesis is divided into two parts. Part I contains three chapters: Analysis of sounds, Treatment of vowels, and Treatment of consonants. Part II contains fifteen chapters: Vocabulary, Word formation, The declension of nouns, The enclitic definitives, Pronouns, Adjectives, Numerals, Verb root, Classification of verb, Conjunctions, Emphatic forms, and Conclusion. Two aspects of this study will be taken up for discussion here: the study of phonology and the study of syntax.

It was observed in §1.3 that it is hard to study the phonology of the language of MKR because of the lack of spoken evidence of that time. Since MKR is a orthographic text, no one knows the extent of correspondence between the sounds of the language of that time on the one hand, and the orthographic representation available to us, on the other. The orthographic conventions of Assamese are heavily influenced by those of Sanskrit. It is evident from the orthography of modern Assamese that there are many letters inherited from Sanskrit which are not pronounced in actual speech in Modern Assamese. During the stages of development through Prakrit and Apabhramsa to
Assamese, the pronunciation of some letters have been lost or otherwise modified. However, the spelling conventions for these altered pronunciations did not correspondingly change. As a result, we still have more letters in modern Assamese as against fewer sounds in actual speech. It is far from clear as to how many letters of the Assamese alphabet found in the manuscripts of MKR were really pronounced in the actual speech of the time. From this perspective, it is difficult to see how any reliable phonological research on the language of MKR can be conducted.

It is evident from the reasons given above that it is not possible to construct sound inventory of fourteenth century Assamese or indeed of MKR on the basis of the orthography of that text. The basis of studies provided in Part I by Saikia (Bora) (1985) is actually orthography (letters), not sounds. Saikia (Bora) contends that the orthography of KR was mainly based on the local pronunciation (1985: 19). However this statement does not seem to be based on any empirical evidence.

As stated earlier, an approximation to the sounds used at the time of MKR can be arrived at only through a process of reconstruction by applying the comparative method among different varieties of present-day Assamese with input from the close cognates of Assamese. In the absence of such efforts, a phonological study based on the orthography does not have any empirical value.

As a linguistic study Saikia (Bora) (1985) appears to cover only phonology and morphology. The discussion of syntax is completely left out in this study. Only form in isolation does not depict a clear picture of any language. The linguistic study of a text without any discussion of the distribution and function of forms loses much of its significance. Most of times, the meaning of forms can be extracted from context only. But in Saikia (Bora) (1985), such factors seem to be totally ignored. In following paragraphs, the syntactic relevance of forms will be highlighted with special reference to some topics described in this work.

Regarding gender, Saikia (Bora) states,

> Assamese does not possess grammatical gender. Only animate objects distinguish the gender in Assamese. The KR also retains this peculiar feature of gender. In KR gender is generally distinguished in two different ways: (i) by the employment of different words which stand either for masculine or feminine and (ii) by adding feminine suffixes. (P. 118-119)

This observation is far from satisfactory for the following reasons:
(a) The statement in the first line contradicts what follows: if indeed Assamese has no grammatical gender, then it is difficult to see how animate objects in Assamese can display a gender distinction. This indicates a failure to show clearly the distinction between grammatical gender and the concept of gender in semantics. This aspect will be further discussed in §6.3.5.

(b) The manner of description is somewhat confusing. Saikia (Bora) (1985: 119) makes some general remarks about Assamese and then proceeds to the discussion of Rāmāyana. However, since no sufficient specimens of the Assamese language of the period before the 14th century are available, it is not clear whether such generalizations can be made with any certainty. While some features of the language can be found in the Caryās, these are not sufficient to serve as a basis for generalizations. Nor is it methodologically appropriate to draw conclusions/make generalizations about the pre-MKR stages of the Assamese language on the basis of modern Assamese as (Saikia (Bora) 1985: 119) has done. Relaxing such restrictions can be justified only in cases where the researcher is able to draw logical conclusions about the kind of change that the data from the earlier stage of the language might have undergone on the basis of some data from a later stage of that language (cf. §7.4.3.1.2). [Indeed, MKR is the first written text of Assamese where so many structural features of the language are found.]

With regard to the category of case, Saikia (Bora) (1985) says,

In Assamese, nouns have seven cases such as nominative, accusative, instrumental, dative, ablative, genitive, and locative.............The Rāmāyana of Mādhav Kandali, a voluminous specimen of early Assamese, retains all the seven cases and most of the case suffixes of Assamese. (P. 122-123)

She further adds that “each case has different case-suffixes to indicate the case-relationship of nouns except ablative. Ablative case is indicated by the addition of post-position para ‘from’, after the genitive -ra.” (P.122)

This description is unsatisfactory from a syntactic point of view. Case is not being considered here from a syntactic dimension. The main function of case suffixes is to mark grammatical relations within the clause. They are used to identify grammatical relations such subject, object, adjunct etc. that occur in a clause. But in Saikia (Bora)
in Rāmāyana: ergative, dative, genitive, and locative. Absolutive case is expressed by zero marking. Ergative, instrumental and locative are expressed by same case marker, though there is a separate marker for the locative in MKR (§7).

Finally, this description is not able to distinguish clearly grammatical case from semantic case. Only an ablative sense is expressed by the addition of genitive -ra 'GEN' with the relator noun parā 'from', not ablative case.

In case of adjectives, Saikia (Bora) (1985) observes,

In the KR adjectives are normally uninflected; but when an adjective denotes feminine gender, then it takes the feminine suffix -i, e.g., kāni, khuri, sundari etc. The KR shows some other adjectives which are not inflected according to gender. For example, ṣāpachu, ktura, ḍūṅgara, bhāla etc. (P. 168)

It was stated earlier that gender is not grammatical in MKR, i.e., it does not show any agreement relation between head words with their dependents. The feminine suffix -i described by Saikia (Bora) above is used to express sex distinction in MKR. From this it cannot be said that MKR has gender, nor can it be generalized in terms of some rule to the effect that adjectives in MKR take the feminine suffix when they denote the feminine gender. The observation made by Saikia (Bora) (1985) is just a coincidence – the part of a derivation process which can be simply considered as an influence of Sanskritism (§6.3.5).

Regarding the classification of verbs in MKR, Saikia (Bora) (1985) states,

The verbs of KR are mainly of two types: Finite and Non-finite....The finite verbs include transitive and intransitive, simple and compound and negative verbs. (P. 202)

This classification does not mention the criterion for classification. Secondly, simple verbs can be both transitive and intransitive, and negative as well. The classification of verbs into finite and non-finite, transitive and intransitive, simple and compound or
complex is usually made on the basis of distinct criteria. However, such criteria are not mentioned here, resulting in an overlapping of the various classes.

Another shortcoming that can be seen in the classification of verb is that the distinction between tense and aspect has not been clearly made in Saikia (Bora) (1985). In fact, there is no reference to Aspect in the discussion. With regard to tense, it is stated,

The tenses of verbs are mainly of two types: Simple and Compound. Simple tense includes present, past, future and perfect while compound tense includes progressive tenses. (*ibid.: 213*)

The perfect and progressive tenses referred to here are actually subcategories of aspect, not tense. MKR distinguishes a binary opposition of aspect: imperfective versus perfective. Imperfective aspect is expressed by the suffix -\((i)\)c/ch 'IPFV', while perfective aspect is unmarked. The progressive aspectual connotation is expressed in MKR through the periphrastic realization of -\((i)\)c/ch 'IPFV', by using another aspectual auxiliary like thāk 'stay', zā 'go' etc. ($\S$4.3).

Saikia (Bora) (1985) further observes, “the perfect tense is formed by the suffix -\((i)\)c-. Examples of perfect tense in the KR are not many” (P. 220). But as we have stated above, -\((i)\)c/ch 'IPFV' is a marker of imperfective aspect, not of perfect tense. However, it is true that -\((i)\)c/ch 'IPFV' serves a dual function in MKR: it expresses an imperfective sense as well as the function of present perfect. This issue will be considered in §4.3.

In case of progressive tense, Saikia (Bora) (1985) states, “the progressive tense includes present, past and future. The present and past progressives take an auxiliary verb āch- while future takes thāk-” (P. 221).

There is some problem with this analysis. Firstly, contrary to the claim expressed in (Saikia (Bora) 1985), the progressive is not a tense. Rather, it relates to aspect and is, indeed one of its sub-categories. Secondly, it is also not clearly defined in this description as to how progressive tense is expressed. Progressive aspect is expressed by the sequence of main verb + ante-aspectizer + aspectual auxiliaries like āch 'be', thāk 'stay', zā 'go' etc. Among these, āch 'be' is a grammatical auxiliary, and it is a
periphrastic realization of the imperfective aspect marker -(i)c/ch 'IPFV' while thak 'stay' and za 'go' are lexical aspectual auxiliaries (§4.3).

There is little discussion on the category of person in Saikia (Bora) (1985). It states,

The KR shows three persons: First, Second and Third. The distinction of person is determined by the suffixes which are called here personal terminations or suffixes. But in the verbal inflections, the distinctions of persons are not strictly maintained in the KR; sometimes Hon. takes inferior terminations and vice versa. (P. 213)

There are several issues with this statement. Firstly, person has not been dealt with from a syntactic dimension. As will be shown later, person markers are used in MKR to show the agreement between verb and subject. They serve to identify grammatical relations in a way that is similar to the case marking system mentioned above. However, this description has not taken this aspect into account. Secondly, the category of person in MKR is too broad and complex to be discussed so briefly in the passing. The category of person will be dealt with in detail in §3.

In terms of its overall organizational plan, certain aspects of Saikia (Bora) (1985) are potentially confusing for the reader. For instance, while case and gender are a part of the chapter entitled The declension of Nouns, the discussion of definitives has been excluded from this chapter. Similarly the chapter on the classification of verbs describes the finite verb and the non-finite verb first, postponing the discussion of tense and mood till after the discussion of the non-finite verb. However, in the organizational plan of grammar, it makes better sense to include the discussion of tense, aspect and mood in the discussion of the finite verb, not to discuss these categories in isolation.

However, in spite of these limitations, it Saikia (Bora) (1985) has to be acknowledged as the first extensive linguistic study of MKR and this work is undoubtedly a valuable contribution to the study of the Assamese language.

Bardalai (2005) has similar shortcomings that are found in Saikia (Bora) (1985). It also contains a discussion of the phonology and morphology of the Rāmāyana, but the data have not been examined from a syntactic point of view.

1.6. Methodology of research

This section presents a general discussion on the corpus based study as a methodological tool for conducting linguistic research (§1.6.2). The process of corpus
1.6.1. Overview

It is normally seen that in traditional linguistic analysis, the discussion of linguistic structures is based on data derived through intuition and introspection. For the present work the researcher has employed a corpus based analysis. The delimited language data of Kandali's Rāmāyana has been entered into a computer and systematically interlinearized and analyzed through wordlist and concordances with the help of linguistic software such as Field Linguist's Toolbox and powerful concordance programme such as CQPweb (Corpus Query Processor). The next section (§1.6.2) introduces the corpus-based analysis as a "methodological basis for pursuing linguistic research" (Leech 1992: 105; cited in Meyer 2002: preface (xi)) in general. The framework of the discussion is mainly based on Meyer (2002), Sinclair (1991), (2003) and (2005), Leech (2005), Hardie (2003). The following sections will indicate the processes that have been employed to extract the grammatical categories from MKR by using the corpus based methodology.

1.6.2. Corpus-based research methodology

Before introducing corpus-based analysis as a methodology, it would be well to provide a definition of a corpus. In general, a corpus is a collection of data that is either written or spoken. According to the Expert Advisory Group on Language Engineering Standards (EAGLES), corpus "can potentially contain any text type, including not only prose, newspapers, as well as poetry, drama, etc., but also word lists, dictionaries, etc." ("Corpus Encoding Standard": http://www.cs.vassar.edu/CES/CESI-0.html; cited in Meyer 2002: preface (xi)). This is a general definition of corpus. However, in linguistic terms, a corpus is defined in a more restricted way than the definition given above. Linguistically, a corpus is "a collection of texts or parts of texts upon which some general linguistic analysis can be conducted" (Meyer 2002: preface (xi)). Though neither definition overtly indicates the size of the corpus, it is understood that a corpus needs to be large in size, since any strong generalization about a linguistic structure is possible only when this generalization is based on a large size of data.
However, it is a very time consuming process for linguists to manually handle a large size of data. For this reason, corpus was produced in computer-readable form in the middle of twentieth century so that linguistic study can be conducted on large amounts of data within less time or effort by using modern technology in the form of analysis software. Regarding the importance of software in a corpus based study, Hardie (forthcoming) comments,

The practice of corpus linguistics is utterly dependent on the availability of computer software capable of effectively processing corpus data by carrying out concordance queries and performing other analytic procedures. Without such software, corpus linguistics may accurately be considered as a ‘pseudo-procedure’, in the words of Abercrombie (1965). (P. 1)

Thus, the definition of corpus given earlier can be modified further by adding the importance of computer technology in creation of corpus as “A linguistic corpus is collection of texts which have been selected and brought together so that language can be studied on the computer” (Wynee 2005: preface).

A corpus can be created for conducting linguistic research on different aspects. For example, the Brown Corpus includes 2,000 word samples of different genres of written English such as press reportage, editorials, government documents, technical writing, and fiction. This type of corpus can be the basis for a systematic study of individual genres of written English and a comparison of the genres. Similarly, the Penn Treebank, which includes a large selection of Dow Jones newswire stories, can be used for conducting research on natural language processing (NLP) The Helsinki Corpus contains various types of written texts from earlier periods of English so that it can be used by historical linguists for the diachronic study of English. The CHILDES Corpus is used by those linguists who are interested in child language acquisition. This corpus contains transcriptions of children speaking in various communicative situations (Meyer 2002: preface (xiii)). Moreover, some linguists use corpora for extracting and describing particular grammatical constructions of any language while others use corpora for writing descriptive grammars of individual languages. Linguists also use corpora for creating dictionaries, developing language pedagogy, and for constructing linguistic theory. In short, corpora can be a basis for undertaking all types of linguistic research. They provide actual instances of speech or writing and hence, the generalisations resulting from corpora are more empirical, more powerful, more real and valid than the
generalisations made from data using intuition and introspection. As a way of pursuing different linguistic research, corpus study has been considered as a methodology of research. To get a good output from corpus-based research, it is necessary to consider some methodological assumptions underlying both the process of corpus creation and the subsequent analysis of a corpus. Three basic processes are found to be involved in a corpus-based study: planning and creating a corpus, annotation, and analysis of corpus. All these three processes are somehow relevant to the present research and thus these will be discussed briefly in the following sections (§1.6.2.1–3).

1.6.2.1. The process of planning and creating a corpus

Generally, the process of planning and creating a corpus is relevant to both spoken and written corpus. However, the process of creating a spoken corpus is not considered here as it is not relevant to the present research. Only the relevant aspects relating to the process of creating the written corpus for present research is described here.

The three steps underlying the process of planning and creating a corpus, i.e., planning the corpus, collecting data and computerizing data are briefly discussed below (§1.6.2.1.1–3).

1.6.2.1.1. Planning a corpus

There are several factors which need to be considered before starting the process of corpus creation. These include the following: the specific purpose for which a corpus being created, the size of corpus, the type of texts to be included in the corpus, the number of samples for each text type, the sources for these texts, the length of each text sample within the corpus, etc. (§1.6.3).

Before determining the size of corpus, it is first necessary to specify the goals of the corpus. If a corpus is being created for analyzing a particular grammatical construction, then this task will be accomplished by using a shorter corpus like ICE corpus which contains only one million words and which was created on the purpose of grammatical analysis of the different regional varieties of English (Meyer 2002: 33). Generally, Corpus can be of different types: Normative corpus, Historical corpus, Monitor corpus, Parallel corpus etc. A normative corpus is designed in such a way that it can represent a close connection to a standard language. A historical corpus is mainly designed to study
the evolution of any language. A monitor corpus is a large corpus which is efficient for
dictionary making. This type of corpus is not static and fixed. It has the capacity of
being updated constantly to include new words and meanings. A parallel corpus is
basically useful for comparative study. Such a corpus gathers texts from more than one
language (Meyer 2002: 15).

After determining the size of the corpus, the next issue is the selection of the texts to
be included in the corpus. There are some criteria to be considered in case of text
selection (Sinclair 2005: 4), (Meyer 2002: 32-52), such as the following:

(i) What type of text will be included in a corpus, i.e., written or spoken text.
(ii) If written, what genres will be included in the corpus, i.e., a book, journal,
letter, notice etc.
(iii) What domain of the texts, whether spoken or written, will be selected for
inclusion in a corpus. For example, if written, whether it is academic or popular,
printed or non-printed etc.; or if spoken, whether it is private or public, scripted
or unscripted etc.
(iv) The length of individual text samples to be included in a corpus, whether such
samples should include text fragments or complete texts (§1.6.3). The selection
of text fragments and complete texts have advantages and disadvantages.
According to Meyer, if someone wants to study grammatical constructions
from any corpora then it is better to select complete texts because “...it is most
natural to study these constructions within the context of a complete text rather
than only part of that text” (2002: 38). Some corpus compilers, however, prefer
to include short samples of more texts from many different speakers and
writers than fewer texts from a smaller number of speakers and writers. In this
effort invested in developing a corpus, broader linguistic representation can be
achieved by focusing on diversity across texts and text types rather than by
focusing on longer samples from within texts.” However, the selection of text
fragments may prevent an analyst from figuring out the infrequent occurrences
of any grammatical construction. In such situations, complete texts would
appear to be a better source.
(v) Determining the number of texts and writers to include in a corpus. Before creating a corpus, it is needed to determine the number of text samples that should be included in a corpus. This would ensure that valid generalizations on a specific aspect can be achieved through the inclusion of the right kind of data into a corpus. In addition, it is very important to select right kind of individuals from whom that data is to be collected.

(vi) The location of the texts.

(viii) The time-frame for selecting texts. The discussion of two types of corpora is relevant here: Synchronic corpora and Diachronic corpora (§1.6.3). While planning the creation of the former type of corpora, attention should be paid to selecting texts created within a relatively narrow time-frame, so that an accurate picture of the contemporary language can be elicited without worrying about the visible effects of language change. Diachronic corpora, on the other hand, is created with texts from the various stages of a language so that the students conducting research on this type of corpora can get an opportunity “to map and compare variant fields or variant paradigms in successive diachronic stages in the past” (Rissanen 1992: 189; cited in Meyer 2002: 46).

(viii) In addition to the criteria given above, there are some sociolinguistic variables which need to be considered before choosing the speakers and writers whose texts are being included in a corpus. These variables are gender, age, level of education, the dialects the individual speaks, the context in which they speak, and the relationship they have with their interlocutors. Among these, the first three variables are relevant for both collecting spoken and written texts while the last three are essential for collecting spoken texts.

After determining the underlying criteria of corpus planning, the next task is started by collecting data. This is considered in the next section.

1.6.2.1.2. Collecting data

The methodology of collecting data varies in terms of spoken texts and written texts. Of these two, the process of collecting speech is much more difficult than the process of collecting written text. However, the data taken for the present research is not spoken
and hence, the methodology of collecting speech will not be considered here. The following considerations relate to the collection of written texts.

The process of collecting written text is much easier than the process of collecting spoken text. Generally, two considerations are involved in the collection of written texts: taking permission regarding the inclusion of copyright texts in a corpus, and whether the texts to be included should be limited to written texts only or extended to include electronically published work.

After planning and collecting texts, the next step is to computerize the texts, discussed in the next section.

1.6.2.1.3. Computerizing data

To computerize written text, the written data can be computerized by retyping it manually or by using an optical scanner and accompanying optical character recognition (OCR) software (Meyer 2002: 55). After typing the text in, the next step is to save the data using an appropriate file format so that it can be supported by any word-processing programme for editing the text.

1.6.2.2. Annotation

Corpus annotation has been described as a process of “adding interpretative linguistic information to a corpus” (Leech 2005: 21). Different kinds of annotations can be carried out on a corpus. These include structural annotation, parts-of-speech (PoS) annotation, grammatical annotation, phonetic annotation, semantic annotation, pragmatic annotation, discourse annotation, stylistic annotation and lexical annotation. Among these, only parts-of-speech annotation is relevant for the purpose of the present research. Such annotation involves a process of assigning a tag to every word indicating its word-class in a corpus (Leech 2005: 22, Meyer 2002: 81).

1.6.2.3. Analyzing a corpus

While analyzing a corpus, three basic methodological aspects need to be kept in mind: the type of linguistic study (i.e. phonological, morphological or syntactic study) to be conducted on the specific corpus, an appropriate corpus for undertaking such study and
the process of extracting the information required for the study from the corpus. Two further aspects are relevant. First, creating data files for keeping record of the information that has been extracted from the corpus and second, determining appropriate statistical tests for analyzing the information in the data files. However, the last one of these has not been covered in the present study. Of these five aspects, the last three can be accomplished by using appropriate software or concordance programme (cf. §1.6.3). Once decisions on these aspects have been taken, the proper process of corpus analysis can be started.

1.6.3. Creating the MKR corpus

For the purpose of the present research, the entire text of Mādhav Kandali’s Rāmāyana was selected for analysis. The Rāmāyana consists of seven kāndas: Adikānda, Ayodhyākānda, Aranyakānda, Kiskindhākānda, Sundarakānda, Lankākānda, and Uttarākānda. However, many scholars are of the view that the Adikānda and Uttarākānda of the Rāmāyana were not originally composed by Mādhav Kandali. It is claimed that the former kānda was written by Mādhavdeva and the latter was composed by Sankardeva (cf. §1.2.1). However, it is not possible to reach any definite conclusion without investigating such claims on the basis of proper empirical linguistic evidence.

It should be clarified here that the corpus of MKR is considered as a synchronic corpus rather than a diachronic corpus. Since this text goes back to the 14th century, MKR might be mistakenly considered a diachronic or historical corpus. However, diachronic or historical corpora are created to study not only language variation in earlier periods but also the changes in the language from one period of time to another. To fulfil this purpose, a historical corpus contains samples of writing from different periods of an earlier stage of a language. As a specimen of an earlier stage of the Assamese language, the corpus of MKR can be considered as a historical corpus. However, from the methodological perspective of corpus construction, the corpus of MKR is a synchronic corpora, because it is a text written in the specific time-frame of the 14th century. The present researcher created this corpus for a synchronic analysis of the language of the time of MKR. In this context, referring to Saussure’s view, Lyons
(1981: 54) observes that synchronic description gets priority over diachronic description. Only through the prior investigation of particular temporal states of a language can we understand the historical development of that language.

As already stated in §1.3, the present researcher has relied on the data collected from the printed version of MKR for conducting research on an aspect of the grammar of MKR. Even though both spoken as well as written corpora are equally important for the descriptive study of a language, using spoken corpora of the 14th century in this study is ruled out because of the unavailability of such data at the present time. Moreover, since the scope of this study was limited to investigating the grammatical categories of MKR, other written texts of this time, even though available, are not included in the corpus.

Every kanda of the version of MKR mentioned above consists of several chapters. A total of 226 chapters are found in the text. In order to create a corpus of the whole text, the process of manually typing the Assamese data with the help of the Unicode supported software Keyman and Unicode font SolaimanLipi were used. For analyzing the text, as mentioned in §1.6.1, the help of software such as the Field Linguist’s Toolbox and the concordance programme CQPweb was taken. In order for both these programmes to apply to the corpus properly, the latter needed to be created by using a Unicode encoding system. For reading concordances supplied by these two programmes, the following methodological sequence provided by Sinclair (2003) has been adopted:

Step 1. Initiate: It is a process of looking words that occur immediately to the right of the hit word and to the left of the hit word and then trying to find out the strongest and regular patterns on the basis of repetitions.

Step 2. Interpret: In this process, some hypotheses are attempted by looking at the repeated words with the assumption that the hypotheses may help to find out a link existing among these words. For instance, the repeated words may be from the same word class, or may all have similar meanings.

Step 3. Consolidate: This involves looking for further evidence for the support of the hypotheses made previously. Evidence is collected at this stage by investigating single occurrences that come close to the hypotheses that have made before, or structures that are different ways of expressing a similar meaning. In addition, hypotheses can be made stronger by observing the adjoining words and the words at a greater distance related to the hits.
Step 4. Report: Writing down the observations obtained from the hypotheses so that there is an explicit, testable version for the future.

Step 5. Recycle: At this stage, along with the words to the right and to the left of the hits, other words from both sides can be investigated by using the facility of sorting. Through this process, anyone can examine the data by sorting more distant words. This process is continued until a repeated pattern is found. Furthermore, if this process retrieves any instances which have not been cited as evidence for at least one hypothesis, it is necessary to examine them to see if there is something that this selection has not sufficiently emphasized.

The following sections (§1.6.3.1 and §1.6.3.2) present brief overviews of Toolbox and CQPweb. The subsequent sections (§1.6.3.3 and §1.6.3.4) focus on two issues: the required formats for inserting the data into both Toolbox and CQPweb, and a comparison of Toolbox and CQPweb. The purpose in providing these sections is to demonstrate the need of the Unicode character encoding system for corpus construction so that this type of corpus can be used for conducting any kind of linguistic research by using linguistic analysis softwares such as Toolbox and concordancers such as CQPweb. This is in recognition of the importance that the Unicode standard has come to occupy in recent global software technology trends. It has been adopted by many industry leaders such an Microsoft, IBM, HP, Apple, Oracle, SAP, Sybase, Unisys, Sun etc. It is required by different modern standards like Java, XML, ECMAscript (Java script), WML etc. and supported by many other operating systems, all modern browsers and many products (http://www.unicode.org/standard/WhatIsUnicode.html).

1.6.3.1. Toolbox

Before turning to the relevance of Toolbox for the present study, a brief idea of the nature and function of Toolbox as a linguistic analysis software will be presented. Different functions of Toolbox will be discussed in the next section (§1.6.3.1.1) with reference to the data from MKR. This and the subsequent section (§1.6.3.1.2) show how Toolbox helps in organizing and analyzing linguistic data.

The Field Linguist’s Toolbox is a computer programme that helps field linguists and anthropologists to integrate various kinds of text data: lexical, cultural, grammatical, etc.
It is especially helpful in organizing and analyzing language data. Based on the data analyzed by this software, linguists can create a collection of annotated texts from a language and write a descriptive grammar and a dictionary of that language. Furthermore, the data analyzed through Toolbox helps software experts to develop taggers and parsers. The programme has flexible options for selecting, sorting, and displaying data. The name Toolbox itself indicates the multiplicity of tools stored together in a box. The word box here indicates a computer programme and within this programme, multiple tools such as database, interlinearization of text, concordancing, wordlist are available. The advantages of Toolbox as an analysis tool for organizing and analyzing the data of the present study will be highlighted below.

1.6.3.1.1. Organizing data

It was stated earlier that Toolbox helps to organize language data. The basic advantage of Toolbox in this regard is that it can deal with large amounts of data. The seven kāṇḍas (chapter) of MKR, containing a total of 226 kāṇḍas in 464 pages are stored together in a single database, named New Rāmāyana in Toolbox. Within this database, different files have been created to facilitate ease of search. While entering the data of Rāmāyana into Toolbox, the following conventions have been used:

1. Distinct files have been created for every kāṇḍa, such as Adikānda, Ayodhyākānda etc. Every 10 chapters are included within a file. Thus, if Adikānda contains a total of 30 chapters, then, three files have been created for organising these chapters, named as Adikānda 1st 10 chapter.txt, 2nd 10 chapter.txt, 3rd 10 chapter.txt, as shown in Figure 1.1. Instead of making groups of ten chapters into a file such as Adikānda 1st 10 chpt.txt, an alternative arrangement would have been to put all the 55 chapters together in a single file, such as Adikānda 55 chpt.txt. However, such a large file would have made it difficult to move around the different records.
Figure 1.1 The organization of chapters in Toolbox

The term *record* needs to be taken up next.

(2) Different records are created for organising the chapters within different files or within a database. The term *record* is used to refer to a complete unit in Toolbox, for instance, to refer to a dictionary entry. In MKR, this term is used to refer to the chapters within a text file as well as to dictionary entries within the dictionary database. Thus, *Adikānda chapter 1* is a record within the file *Adikānda 1st 10 chapter.txt*. Similarly, the lexeme *kara* ‘hand’ is a record within the database called *dictionary.txt*. The name of the record is inserted in the place indicated as *id* name in text window of Toolbox. Figure 1.2 and 1.3 given below are examples of different records under the text database.

Similarly, Figure 1.4 and 1.5 are instances of different records under the dictionary database.
Name of file

Name of file

Name of file

Name of record

Name of Kanda

Number of Chapter

Number of Verse

Figure 1.2 Records in a text window in Toolbox

Figure 1.3 Different records in text window
(3) Using reference numbers to indicate the number of lines in a text.

This is a very useful system for organising text, especially, for keeping records of example sentences. MKR was composed in sloka (verse) form, with each sloka containing two lines. The reference number of every sloka in MKR is given serially side by side. Toolbox has a separate field reference (\texttt{ref}) for keeping the record of verse numbers, as shown in Figure 1.3. A field is used to refer to a particular type of data in a record in Toolbox. For example, in the dictionary entry for \textit{kara} ‘hand’ mentioned above, the first field is Lexeme (\texttt{lx}), for instance, \textit{kara}. The second field is Parts of speech (\texttt{ps}), for instance \textit{N}.

Each field has the Field Name on the left of the vertical line and the data on the right.
I.6.3.I.2. Analyzing data

After organizing the data files into Toolbox, the next task is the interlinearization of data. Interlinearization is a process of annotating texts in Toolbox. This process helps to analyze the text by breaking the words into morphemes, and by glossing each morpheme in the corresponding space in the next line. To do this task, Toolbox uses the dictionary to parse word into morphemes, and to find the gloss and part of speech of each morpheme. Without using the dictionary, the process of interlinearization cannot be undertaken in Toolbox.

Four types of annotations have been used to describe the data of MKR, as shown in Figure 1.6: Morpheme-break for each word, gloss in English, the part of speech for each word and gloss in the regional language. In Figure 1.6, \\mb stands for the morpheme break line; \\ge stands for the gloss-in-English line; \\ps stands for the part of speech line; and \\gr stands for gloss-in-regional-language line.

Figure 1.6 The model of interlinearized texts in Toolbox
Apart from the advantage of annotating texts, interlinearization is also a useful process for studying the morphology and syntax of a language. Two different tools are available here for this purpose: Word list and Concordance. A wordlist basically helps to analyze the morphology. Toolbox offers the option of creating two wordlists in two windows, sorted from opposite ends. This helps the investigator to see all the words that begin in the same letter and all the words that end in the same letter. Moreover, the process of creating wordlists displays a list of occurrences of the same root or prefix together in one window, and a list of occurrences of the same suffix together in the other window. The two wordlists created for analyzing the data in MKR are shown in Figure 1.7.

**Figure 1.7 The two wordlists created in Toolbox**

It can be seen that the element अमासाक ‘us’ is highlighted in both wordlists. Both wordlists help to identify this word as a sequence of two morphemes, i.e., अमात ‘we’ and ताक ‘you’, instead of a single word. The evidence is that अमात is not only found to occur with ताक, but also with other elements such as सार, थेरा, दा, टा etc. displayed in Word list 1. Likewise, ताक not only occurs with अमात, but is a common element for both अमासाक ‘us’ and तोमासाक ‘yours’ as displayed in Word list 2. From this point of view, अमात can be regarded as a base, because it is the common element in all the occurrences in Word list 1. ताक, on the other hand, appears to be a suffix, because it occurs in the final position, and is a common element of both the words अमासाक and तोमासाक. However, further data in Word list 1 reveals that ताक cannot be considered as a single element.
Rather, it consists of the two morphemes -sā and -ka. The following element of āmāsāka in Word list 1 is āmāsāra which means ‘our’. -ra in this occurrence functions as a genitive case marker. This marker is not only found with this pronoun, but also with all type of nouns and pronouns. This is evident from the pronoun āmāra ‘our’ in Word list 1. -ka in āmāsāka occurs as a dative case marker in MKR. It occurs with all types of nouns and pronouns as a dative marker. āmāka ‘us’ in Word list 1 and ghōrāka ‘horse’ in Word list 2 can be cited as evidence for our contention. The distribution of -ka in āmāsāka is the same as that of -ka in āmāka. Similarly, the distribution of -ra in āmāsāra and āmāra is the same. The -ka in the first pair marks the grammatical relation of object, while the -ra in the second pair marks the head-modifier relationship.

It is reasonably clear that -ka and -ra are dative and genitive markers. On the other hand sāka cannot be considered to be a single morpheme. Rather, the element -sā within sāka would seem to indicate something else. In order to determine the function of this form, the concordancing tool can be used. It was stated earlier that a concordance helps us to see the distribution of forms in context. Using this tool, the distribution of any form containing -sā can be examined, as shown in Figure 1.8. Here, the form tomāsāka is taken as the target form. Except in Laṅkā – 07.036 (Laṅkākānda – Chapter 7, Verse 36), this form is used for plural reference in all its occurrences. From this point of view, -sā can be considered as a plural marker.

Figure 1.8 Concordance list in Toolbox for the word tomāsāka
Apart from identifying the function of forms, a concordance also helps an investigator to find out the immediate constituents of elements i.e. to find out what element goes with what other element in syntactic terms. For example, the postposition *lägi* in MKR governs the case marker -*ka*, and indicates ‘motion towards’, ‘extent’, ‘duration’, and ‘purpose’. However, *lägi* is also found to occur without the governing -*ka* marker, even though this type of structure is not as frequent as the former structure.

Figure 1.9 illustrates.

**Figure 1.9 Concordance list in Toolbox for the word *lägi***

It may be noted here that the method of concordance is not only useful for searching word forms, but also for searching annotations such as grammatical tags, as shown in Figure 1.10.

**Figure 1.10 Concordance list created in Toolbox for the parts of speech N (noun)**
The kind of concordance list shown in Figure 1.10 helps to extract the information of associated grammatical categories of particular word-classes. Thus, it can reveal what type of suffixes or prefixes are found to occur with nouns, or what kinds of words are allowed to occur with particular word-classes. However, it may be noted that for searching annotations, the data files need to be annotated first. Without such an annotating process in place, the kind of grammatical information indicated above cannot be extracted.

The foregoing discussion was intended as an illustration of how word lists and concordances of Toolbox can help linguists to investigate and analyze the morphology and syntax of a language. There are other tools available within this programme for help with dictionary making. Figure 1.11 displays the main Toolbox window containing a text window, a dictionary window and two wordlist windows. Along with the concordance window, it is possible to work on five windows in Toolbox at the same time.

Figure 1.11 Different windows displayed on Toolbox
1.6.3.2. **CQPweb**

Like Toolbox, CQPweb is also relevant to goals of present research, i.e., to extract information on the grammatical categories from the data of MKR. The following discussion introduces the features of CQPweb before dealing with how some of these features have been used for the extraction of grammatical categories found in MKR.

CQPweb is a web-based corpus analysis tool designed by Andrew Hardie of Lancaster University. The purpose of developing this programme is to improve the two important requirements for corpus analysis software: *usability* and *power*. The concept of usability is defined on the basis of user-friendliness. The concept of power is defined on the basis of two criteria: first, a tool can be considered as powerful if it can query a large or very large corpus containing tens or hundreds millions words within a short time span. It is true that the speed of any query depends partly on the computer hardware being used. However, if the datasets being queried are not indexed\(^3\), the programme will not be able to achieve the highest speed for a searching query on large datasets despite high-tech (high-end) hardware. Second, a programme can be considered as powerful, if it can be used for complex and sophisticated queries, i.e., it can be used to search annotations such as grammatical tags or text headers via an abstract data model. From these perspectives, CQPweb is extremely powerful, because, it can query a very large corpus quickly, and it has the type of query language that can be used for complex and sophisticated queries. The most powerful feature of CQPweb is that it deals with various types of corpora.

CQPweb provides a variety of tools by which any kind of linguistic information can be extracted. Some of these important tools are thinning, *extraction of collocations*, *creation of distribution tables and charts*, *categorising*, *frequency break-down list*, and *sorting*. Of these tools, only the facility of extracting frequency break-down lists and sorting have been used for the present purpose. These two tools can be defined in the following way:

---

\(^3\) The information on CQPweb discussed here is taken from Hardie (forthcoming). This information seeks to briefly introduce the programme first before discussing its relevance for the analysis of the data for the present study.

\(^4\) A corpus is indexed when the data structure of a particular corpus permits the query programme to locate matches for a query without searching sequentially through all the texts of the corpus.
(1) Frequency break-down list: The purpose of this list is to show the frequency of a particular form that occurs as a hit word. It also helps to show the primary annotation of a specific hit word. The frequency list of a hit word can be achieved by clicking on that hit. The process of clicking will be able to immediately produce a query thinned to just the subset of hits that match that word.

(2) Sorting: This is a most essential tool for linguistic analysis. Queries can be sorted on the hit word or a selected position before or after the hit word; the sort can be by word or by the primary annotation. In addition, sorting has the facility of thinning a query. The list obtained from this process of thinning helps to specify a pattern that the word/annotation at the sort position must match to be shown.

While CQPweb offers much more features for corpus analysis than those outlined here, all those features do not directly bear on present research.

It was mentioned earlier that CQPweb can work with a variety of corpora. It has a powerful search engine which helps to search a query over hundred million words within a short time span. The main page of CQPweb provides the information about how many corpora it is handling currently. Mādhav Kandali's Rāmāyana is one of the many corpora (indicated in a box in red) occupying a place in CQPweb as shown in Figure 1.12. The preparatory activities undertaken by the present research to include the MKR data in CQPweb will be discussed in the next section (§1.6.3.3.).
Figure 1.12  A display of the various corpora found in CQPweb

By clicking on the required corpus, any analyst can go to the main menu from where various operations on the corpus can be started. Figure 1.13 shows the main menu of CQPweb. Any query can be searched by clicking on the Word lookup (indicated in a box in blue) menu on the left. The word lookup menu is already highlighted in Figure 1.13. The query expression needs to be typed into the window which says Enter the word-form you want to look up (indicated in a box in green). Next, one of the options
has to be chosen from *show only words*... (indicated in a box in dark blue) as shown in Figure 1.13. The option *starting with* (indicated in a box in dark blue) helps to look at the distribution of prefixes. Another option, *ending with* (indicated in a box in yellow), helps to look at the distribution of suffixes. A further option, *containing* (indicated in a box in red), is used to serve both purposes. *matching exactly* (indicated in a box in black) is an option which enables an analyst to make a query that matches exactly with the word or expression in the data. After choosing the relevant option, the analyst has to click on the *lookup* (indicated in a box in black) button. This process helps anybody to search for a particular word or expression in a database of over hundred million words within a very little time.

![Figure 1.13 The process of searching a query on CQPweb](image)

On using the Word lookup menu to look for the occurrence of the suffix -ante in MKR as shown in Figure 1.13, the query processor of CQPweb immediately brings up a frequency list of this suffix. This list is displayed in Figure 1.14.
The display indicates that there are 182 words available in MKR which contain the form -क्ष (ante) at the end of the word. However, -क्ष does not function as a suffix in all those words. Only the words shown in violet contain -क्ष (the others are just single words, not a combination of two morphemes). Out of 182 types, only 20 types are shown in Figure 1.14. The common feature of all the words shown in violet is that they are all verb.
forms which share -\textit{a\^s} as a common element. It is thus recognized as a suffix. Even then, it is still not known what type of suffix it is and what is its function. To find out about the function of -\textit{a\^s}, we can check the distribution of all the words by clicking on each word in turn. CQPweb brings out a menu which shows the distribution of all words ending in -\textit{a\^s}. Thus, if we first click on the word \textit{k\^ante}, a list of occurrences of this word will come up as shown in Figure 1.15 below.

![Figure 1.15 The concordance list of \textit{k\^ante} in CQPweb](image)

The different functions of -\textit{a\^s} can be extracted from this list. If some information is not available in this list, CQPweb offers a further way to look at the data by using the facility of \textit{sorting} by which anybody can investigate the function of the particular hit with the help of its collocates. CQPweb has the facility of sorting till five right and five
left to the hits. Figure 1.16 given below is an example of the sort of the words occupying the slot immediately to the right of the hit ‘कम्पन्न’ (indicated in a box in blue). The list shown in Figure 1.15 helps to introduce -आम as a non-finite marker. In addition, this list also indicates that though this marker functions as a non-finite marker as a whole, its distribution as a non-finite marker is not the same in all contexts. This is evident from Figure 1.16.

**Figure 1.16 The concordance list of कंदन्ते (kāndante) with one right sorting**

Figure 1.16 indicates that the structure of 1, 4, and 10 are alike. In these three occurrences, the hit word occurs with the auxiliaries अच्छा (āchaya) in 1 (the element that occurs in the one right position of the hit word), अच्छ (āchanta) in 4, and थाकिला (thākilā) in 10. All of them carry an aspectual connotation of continuation. These occurrences suggest that -आम functions as ante-aspectizer. However, in other occurrences, it expresses an adverbial function, as in 13 and 17. In these examples, the verb form containing -आम functions as an adverb of reason. The distribution of all the forms in violet in Figure 1.14 serves to show that the verbs with -आम are non-finite forms. Any
investigation of their other functions as non-finite forms is beyond the scope of the present discussion which was intended to give an idea of the methodology of extracting grammatical information from corpora and the relevance of corpus analysis for linguistic study.

1.6.3.3. The format for inserting MKR data into Toolbox and CQPweb

This section seeks to explain the reason for using the Unicode standard in the process of corpus creation of MKR. It also sets out the process that was followed for including MKR in CQPweb.

It was mentioned earlier that the data of MKR needed to be typed in the Unicode format for entering into both Toolbox and CQPweb. Over the last decades, the Unicode has emerged as a unique character encoding system which can be used to represent the characters from a very large number of writing systems. This system provides a unique number for every character existing in the languages of the world so as to match with any platform, any programme and any language. As a character encoding system, Unicode helps to interchange electronic data between different computer systems in different countries without any confusion or mismatch. This is the reason why the Unicode format has been used to create the corpus of MKR. As a result of this exercise, the data can now be made available on the internet, thereby creating opportunities for other researchers for conducting further research on both the linguistic and literary aspects of this work.

In addition, Unicode is of help in putting together texts from different writing systems into a single document. In other words, it provides a facility for working with multilingual documents or for working with multilingual corpuses from within the same web browser. However, for the purpose of transmission of electronic data over a network connection, out of the three Unicode transformation standards (UTF)\(^5\) UTF-16 (Unicode itself), UTF-8, and UTF-32, UTF-8 is considered as the most popular way of exchanging Unicode data between entities in a heterogeneous environment, and as a way of avoiding the 'textual Tower of Babel'\(^6\) that may be seen in the process of

---

\(^5\) For details of what UTF is and why it is considered as a more popular way of exchanging Unicode data among different computer systems internationally, see McEnery and Xiao (2005).

\(^6\) This expression is adopted from McEnery and Xiao (2005).
interchanging Unicode data. For putting the corpus of MKR in CQPweb, the present researcher had to use UTF-8 encoding. For this purpose, the texts needed to be typed into the notepad as text documents and saved as UTF-8 encoding.

1.6.3.4. Toolbox, CQPweb and Interoperability

This section seeks to highlight the complementary nature of both Toolbox and CQPweb along with the possibility of developing the process of interoperability between these two programmes.

While discussing the advantages of Toolbox, it was stated that it can deal with large amounts of data. However, because of its ability to query over the hundred millions of words, CQPweb is more powerful than Toolbox. Moreover, CQPweb's ability to query different parts of words is not matched by Toolbox. Toolbox does permit the query of various parts of words, but only after the annotation of the data files is done. However, it must be admitted that the purpose of designing these two programmes is different. Toolbox is basically designed to help field linguists when they work in the field. It is best for working with smaller amounts of data in comparison to CQPweb. As such, it offers various tools for an in-depth and micro-level analysis of language data. CQPweb, on the other hand, is designed for the ability to carry out a query over millions of words within a short time span. Due to this focus on power, the analysis of language data on CQPweb is relatively shallow in nature. In short, Toolbox helps to look at the language data from a \textit{worm's eye view}, while CQPweb helps to look at the data from a \textit{bird's eye view}. The latter offers a range of tools for a macro-level analysis of language data.

Using CQPweb, linguists can examine large amounts of data from a broader perspective, while Toolbox helps to select small part of them and allows linguists to conduct their research by segmenting the data into relevant smaller bits. Observing the prospect of interoperability between these two programmes, the designer of CQPweb, Andrew Hardie, decided to develop a system where the interlinearized texts done on Toolbox and the data available in CQPweb can be operated together from same CQPweb page and server model.
1.7. Theoretical framework

The analysis and description of the language material of Kandali’s Rāmāyana has been carried out within the framework of descriptive linguistics as elaborated by its proponents from time to time. During this process, concepts and insights from general linguistics have been brought in as and when the need arose, and as driven by the research questions. General and descriptive, both subfields of linguistics are complementary in nature. Both are dependent on each other explicitly and implicitly. The interdependency having between these two fields can be explained in terms of the following observation of Lyons (1981: 34): “General linguistics supplies the concepts and categories in terms of which particular languages are to be analyzed; descriptive linguistics, in its turn, provides the data which confirm or refute the propositions and theories put forward in general linguistics.” Elaborating the interdependence between the two fields, Lyons further observes,

The general linguist might formulate the hypothesis that all languages have nouns and verbs. The descriptive linguist might refute this with empirical evidence that there is at least one language in the description of which the distinction between nouns and verbs cannot be established. But in order to refute, or confirm, the hypothesis the descriptive linguist must operate with some concepts of ‘noun’ and ‘verb’ which have been supplied to him by the general linguist. (1981: 34)

For example, the general linguist might hypothesize that all the languages across the world exhibit five patterns for case marking (§7.1), and suggest some strategies for identifying these patterns. The descriptive linguist, in his or her turn, can confirm or refute this hypothesis with empirical research on some particular language by using the strategies provided by the general linguist. A general linguist might supply some criteria for identifying word classes in general terms (§2.2). In that case, the task of the descriptive linguist would be to use those criteria to find out the word classes of a particular language. Since the present work is descriptively oriented, the present researcher has adopted many concepts and ideas from general linguistics with the goal of applying those common concepts and ideas on the data of MKR for uncovering the grammatical features of Assamese relevant to the study.

Furthermore, given a choice between emphasis on language data and linguistic theory, the present work has given a higher priority to the task of accounting for and
explaining language data. The primary goal of the present research is not to focus effort on developing some theory-centric approach. Only such theoretical resource and perspective that is relevant to adequately describe the language material has been taken into consideration.

1.8. Issues encountered while analyzing the data

Several issues were encountered while the analysis of the data of MKR was being carried out. It is the purpose of this section to briefly set out the nature of these issues and the problems they created while examining the grammatical patterns for categories such as case (§7), person (§3) etc.

(1) Orthographic issues

Inconsistencies in the use of orthographic conventions have been one of the major problems encountered in the work. Different spellings have been used to write the same word. Some of the frequent irregularities found in spelling usage of MKR are the following: the alternate use of some letters within the same word as $i (இ)$ and $i (இ)$; $u (ǚ)$ and $u ( %[ू) ; th (ฑ) and th (ฑ) ; dh (ฑ) and dh (ฑ) ; j (ஜ) and jh (நி) ; n (னி) and n (நி). Examples of each pair are given below as illustrations. The target alphabet is highlighted in bold.

\[ \begin{align*}
& i (இ) \text{ nirikṣi (निरिक्षी) (kis, Ch. 24, V. 58) and } i (இ) \text{ nirikṣi (निरिक्षी) (Lan, Ch. 46, V. 46)} \\
& u (ǚ) \text{ durara (दूररा) (Su, Ch. 16, V. 84) and } u ( %[ू) \text{ durara (दूररा) (Su, Ch. 7, V. 39)} \\
& t (ฑ) \text{ sito (सित) (Ay, Ch. 2, V. 14) and } t (ฑ) \text{ sito (सित) (Ay, Ch. 7, V. 33)} \\
& th (ฑ) \text{ thāna (थान) (Ay, Ch. 3, V. 28) and } th (ฑ) \text{ thāna (थान) (Ar, Ch. 6, V. 95)} \\
& dh (ฑ) \text{ dhopa (धोपा) (Lan, Ch. 15, V. 39) and } dh (ฑ) \text{ dhopa (धोपा) (Kis, Ch. 5, V. 13)} \\
& j (ஜ) \text{ jāka (जाका) (Ay, Ch. 1, V. 14) and } jh (நி) \text{ jāka (जाका) (Ar, Ch. 7, V. 15)} \\
& n (னி) \text{ tūne (तूने) (Lan, Ch. 48, V. 255) and } n (னி) \text{ tūpe (तूपे) (Kis, Ch. 18, V. 39)} \\
\end{align*} \]

Such orthographic irregularities are found not only in MKR, but in other old Assamese texts as well. While commenting on the orthographic inconsistencies of MKR, Medhi (1913) observes, “the Orthography of KR is very confusing. This, however, is
not peculiar only to the KR as most of the old Assamese texts show such orthographical irregularities” (Preface of Prahlād Charitra, cited in Saikia (Bora) (1985: 17).

According to Sharma, this type of orthographical corruption is not only found in old Assamese texts, but in the Sanskrit texts of the ancient inscriptions of Assam, during the period from 5th century AD to 12th century AD (1978: 0.24-0.25 (Introduction)). He further says that if those Sanskrit texts are examined carefully, it can be seen that even the orthographic corruptions found in ancient inscriptions of Assam display a pattern which may be due to the influence of local Prakrit on Sanskrit texts (§1.1). There may be various reasons for the irregularities in the use of orthography in MKR. Some of these reasons cited by Saikia (Bora) (1985) are given below. Further reasons will be discussed in a later chapter (§3.3):

(i) The scribes of MKR might have been familiar with different varieties of Assamese scripts and hence, they might have been confused in using the appropriate script, especially in the case of homophonic letters. This may have led them to use two or three different spellings for the same word. (What Saikia (Bora) refers to as homophonic should be more properly considered as homographic, since the data of MKR is written, not spoken text).

(ii) It is difficult to find the original hand-written copy of the Rāmāyana composed by Mādhav Kandali.

(iii) There was neither a dictionary nor a grammar of the Assamese language at the time when the MKR was written. For this reason, the scribes could not maintain the standards in spelling.

Orthographic irregularities of this kind prevent us from getting a clear idea of the orthographic system of the language of the time of MKR. Furthermore, the absence of consistent orthography creates further problems in attempting any phonological analysis on the language data.

(2) Irregularities in form-function alignment:

Two kinds of irregularities are observed in this regard: (i) the same form used for different functions, and (ii) Different forms used for the same function.

The first type of irregularity is frequently found in the use of case markers (cf. §7). For example, the -e marker is used for variety of functions such as marking subjects,
location and instrument. In some cases, it is used to mark the direction and adverbs of manner, time etc. Similarly, the -e and -ka markers are used to mark both direct and indirect objects as well as to mark direction, duration etc. Person markers exhibit similar irregularities. For example, the person agreement marker -e occurs with both second person and third person subjects without showing any underlying regularity (cf. §3). (4) and (5) below illustrate irregularities in person marking.

(4) tai ayodhya napāibe
tai ayodhya na-pā-ib-e
you.INF Ayodhya NEG-get-FUT-2INF
'You will not reach Ayodhya' (Laṅ, Ch. 34, V. 13)

(5) tebe kiya āmāka daṇḍibē nṛpabara
tobe kiya āmā-ka daṇḍ-ib-e nṛpabara
the why we-DAT punish-FUT-3 king
'Then why will the king give us punishment?'
(Kis, Ch. 24, V. 13)

In both examples, the verbs take the same agreement marker even though the subjects are different in terms of the category of person. In (4), the subject is tai ‘you’, but in (5), the subject is nṛpabara ‘king’.

Different forms expressing the same function can be frequently seen in instances of person agreement (§3). Two or three forms are used to mark the same function, without any further distinction among those forms. All forms are treated equally for expressing the same function, as in the case of markers such as -ā, -āhā and -e. These three forms are used as second person familiar agreement markers, while -sa, -ha, -a are used as second person inferior agreement markers. Likewise, both -o and -ta, are used to mark location.

Such issues of irregularity are discussed under the heading of syncretism in subsequent chapters. On the matter of using glosses for instances where the same form expresses different functions, the labelling has been done according to the function of
the given form. For example, if -ka is found to be used in three functions i.e. for object marking, direction marking and distance marking, then the gloss has been used accordingly.

Further, it has been observed that consistency has not been maintained in the use of certain forms, as in the case of the -e marker. This marker is obligatorily present with transitive subjects but optionally so with intransitive subjects. It functions as a case marker to express one of the core functions of a clause in MKR, i.e., the subject function. However, the function of the -e marker within a case marking morphology is an important question for investigation. It has been observed that the majority of its occurrences follow a regular order. The rest of the occurrences that do not follow the regular order may be due to the poetical nature of the text. From the data it appears that the presence or absence of some forms is, in fact, due to the need for adjusting the rhyming pattern of the text (cf. §7). In order to extract the regular order from such a marked order, certain conclusions have been drawn from the typological nature (i.e. the prototypical behaviour) of the language group. An example can be cited to illustrate this point. Ergativity is a typological feature of Indo-Aryan languages. While trying to determine the function of -e in MKR, this generalisation can be considered as a hypothesis. It is generally assumed that a group of languages from one language family behaves in a similar way. From this point of view, it can therefore be assumed that Assamese exhibits ergativity in its case marking morphology, since most of its close cognates share this feature. However, if the data available to us shows that Assamese does not actually exhibit an ergative case marking morphology, but a nominative one, then other factors for this divergent behaviour of Assamese have to be found. In this regard, one major factor may be the contact situation within the same linguistic area. If the languages of the area belong to more than one language family, there is a likelihood of the transfer of some structural features from the different languages of that linguistic area to one another. In fact, it is an important issue to investigate if one or two languages of one language family behave differently from the rest of the members of the group. In that case, such differences may be due to the language contact situation within the linguistic area. However, in this case, the dominant group should be figured out first, which is likely to be the language possessing written literature. As far as the function of
the marker -e in MKR is concerned, the available data clearly suggests that its function is ergative, not nominative. It is true that Assamese has been surrounded by the languages of the Tibeto-Burman language family from the very beginning. Therefore, it is not surprising that Assamese has come to share some structural features with these languages. However, information on the political and linguistic situation of the 13th and 14th century found in the history of early Kāmrupa tells us that the language of that time was predominantly Assamese. The history of old Kāmrupa also says that the old Kāmrupa not only contained the geographical boundary of present-day Assam, but also some parts of contemporary West-Bengal and Bihār. Hence, from this point of view, it can be assumed that in spite of the areal influence of Tibeto-Burman due to contiguity, Assamese had close connections with other NIA languages. It is reasonable to assume that the ergative -e marker of Assamese is one of the genetically inherited features shared with its close cognates unaffected by the features of surrounding Tibeto-Burman languages.

It may be noted here that some forms, mostly in the person paradigm, have been used only once or twice in entire text. Even though such forms may be orthographically irregular, these have not been considered in the present discussion since they do not seem to carry any significance.

1.9. Structure of the present work

This section is mainly about the structure of the present work, i.e., how many chapters are there in the present work and what relationship is found among the chapters. While organising these chapters, an attempt has been made to follow a certain pattern. Further remarks on the rationale of the organisation of chapters will be presented in Chapter Two (§2) after certain operational notions (such as Inherent/Relational) related to research topic have been discussed.

The present work is divided into eight chapters including this introductory chapter (§1). Chapter Two (§2) introduces the major word-classes and their associated grammatical categories from both the perspectives of general linguistics and MKR respectively. This chapter also provides the underlying order of presentation plan for the latter chapters.
The next two chapters (§3 and §4) focus on the grammatical categories of the word-class verb. Chapter Three discusses the agreement category for the verb, i.e., the grammatical category of person from both the perspectives of affixations to the verb, viz., personal suffixes and from the perspective of independent stems, viz., personal pronouns. Chapter Four deals with the inherent categories for the verb, and discusses tense, aspect, and mood. The discussion of negatives and causatives are also included in this chapter. Chapter Five (§5) provides an elaborative discussion of non-finite markers available in MKR from the perspectives of both forms and functions.

The grammatical categories for nouns are discussed in Chapters Six and Seven (§6 and §7). Chapter Six provides a description of the inherent categories for the noun, and covers the discussion of the classifiers, personal deixis, numerals and other modifiers of the noun element. Chapter Seven focus on the relational category for the noun, i.e., the case marking system in MKR. Chapter Eight (§8) undertakes a brief critical analysis of the work done in the previous chapters along with the future prospects of the present study.