Construction and annotation of a corpus of contemporary Nepali

Yogendra P. Yadava,[1] Andrew Hardie,[2] Ram Raj Lohani,[1] Bhim N. Regmi,[1] Srishtee Gurung,[3] Amar Gurung,[3] Tony McEnery,[2] Jens Allwood[4] and Pat Hall[3]

Abstract

In this paper, we describe the construction of the 14-million-word Nepali National Corpus (NNC). This corpus includes both spoken and written data, the latter incorporating a Nepali match for FLOB and a broader collection of text. Additional resources within the NNC include parallel data (English–Nepali and Nepali–English) and a speech corpus. The NNC is encoded as Unicode text and marked up in CES-compatible XML. The wholecorpus is also annotated with part-of-speech tags. We describe the process of devising a tagset and retraining tagger software for the Nepali language, for which there were no existing corpus resources. Finally, we explore some present and future applications of the corpus, including lexicography, NLP, and grammatical research.

1. Introduction

Nepali is an Indo-Aryan language spoken by approximately 45 million people in Nepal, where it is the language of government and the medium of much education, and also in neighbouring countries (India, Bhutan and Myanmar). It serves as the lingua franca of an extremely multilingual part of the world: more than ninety languages are spoken within Nepal.[5] Nepali is written in the Devanagari alphabet and has a written tradition extending back to the twelfth century. Until recently, however, there has been no work on corpus development or corpus analysis for the Nepali language. Indeed, Nepali has been largely excluded from access to information and communication technology in general.

[1] Central Department of Linguistics, Tribhuvan University, Kirtipur, Kathmandu, Nepal
[2] Department of Linguistics and English Language, Bowland College, Lancaster University, Lancaster, LA1 4YT, United Kingdom
Correspondence to: Andrew Hardie, *e-mail*: a.hardie@lancaster.ac.uk
[3] Madan Puraskār Pustakālaya, PO Box 42, Shreedurbar Tole, Patan Dhoka, Lalitpur, Nepal
[4] Department of Linguistics, University of Göteborg, S-412 82 Göteborg, Sweden
[5] According to the 2001 census of Nepal.

DOI: 10.3366/E1749503208000166
Corpora Vol. 3 (2): 213–225

This issue has recently been addressed by the *Nelralec*[6] project, known in Nepali as *Bhasha Sanchar* (literally 'language communication'), undertaken by a consortium of Nepali and European partners including the Open University, UK; Madan Puraskar Pustakalaya, Nepal; Tribhuvan University, Nepal; ELRA; the University of Göteborg, Sweden; and Lancaster University, UK. A variety of Nepali language technology support projects were undertaken within Nelralec, including software localisation and font development. In this paper, however, we report on the consortium's work towards the development of the *Nepali National Corpus* (NNC), which was completed in late 2007.

In Section 2, we will explain the design of the various components of the NNC, elaborating on the problematic issues that we faced in assembling the corpus texts. We will also outline the applications to which the corpus data has been applied to date. Section 3 describes the annotation of the corpus – specifically, the development of a part-of-speech annotation scheme and the training of a Nepali tagger. Finally, Section 4 outlines some future directions of research involving the newly-available Nepali corpus data.

2. Corpus construction

The Nepali National Corpus was conceived as a compendium of different types of corpora, each one incorporating a wide range of Nepali texts. It comprises two separate written corpora, a spoken corpus, a collection of Nepali–English and English–Nepali parallel data, and a speech corpus. In this section, we outline the design of each of these components; the overall composition of the corpus is outlined in Table 1.

The written part of the NNC was designed according to a 'corepenumbra' model, which, as far as we know, is unique to the NNC. In short, one part of the corpus was carefully designed to follow a standard sampling frame, to ensure comparability with similar corpora in other languages, but was, as a result, necessarily limited in size. The other part of the corpus, by contrast, had a much less specific design and sampling frame, allowing us to be less selective about the texts that were incorporated, and was, therefore, able to be made much larger. This model of corpus design combines the advantages of a small-corpus approach, where great attention is paid to representativeness and balance, with the opposite advantage simply of having a very large amount of data, and thus increasing the absolute number of examples that may be found for less common words or constructions.

In constructing the core part of the corpus, we aimed as far as possible to follow the sampling frame of the FLOB and Frown corpora (described in detail

[6] *Nepali Language Resources and Localization for Education and Communication*. The project was funded by the EU Asia IT&C programme, reference number ASIE/2004/091–777.

NNC Component	*Contents*	*Size in words (approx)*
Core Sample	Written texts sampled as a Nepali match for FLOB and Frown	800,000
General Collection	Written texts opportunistically collected, including text from the Web	13,000,000
Parallel data	Written texts with translations Nepali-English and English-Nepali	4,000,000
Spoken corpus	Spoken texts	260,000
Speech corpus	Audio recordings of sentences for use in text-to-speech applications	6,000

Table 1: Components of the Nepali National Corpus

by Hundt *et al*., 1998; Hundt *et al*., 1999). Briefly, this sampling frame selects 500 texts, each of 2,000 words, from fifteen genres. All texts were published in 1991 (this being the sampling year of FLOB and Frown, and allowing direct comparability). For our 'Core Sample' (NNC–CS), we aimed to provide a Nepali match for FLOB and Frown, following the example of McEnery and Xiao (2004), who describe a Mandarin corpus that is comparable in design to these English corpora. However, the NNC–CS is the first example of a corpus in a South Asian language built according to this scheme (the Kolhapur Corpus follows a similar sampling frame, but for Indian *English* only; see Shastri, 1986).

Some adaptations of the FLOB sampling frame needed to be made. Not all the genres that can be identified in English actually exist in Nepali. For example, the *Western and adventure fiction* genre represented in FLOB and Frown has no clear counterpart in Nepali; on the other hand, no examples of *science fiction* could be found within the required timeframe. For this reason, a single fiction genre (labelled 'S') was used to match all the variegated fictional sub-genres distinguished in the FLOB sampling frame. However, the *major* genre distinctions (e.g., press reportage, academic prose, fiction) involved in the sampling frame could all be found for Nepali. Only 398 of the target 500 texts are included in the current release of the NNC–CS. This is due to the time period (1991–2) being sampled; at this time, the quantity of publication in Nepali was relatively restricted. We hope to amend this in future releases. All of the texts were keyboarded. This was

possible because of the relatively small size of the target corpus (one million words), and necessary because of the age of the texts.

A selection of 1,880 sentences (6,053 words) from the Core Sample formed the basis of another part of the NNC, the speech corpus. The choice of sentences was made randomly, with subsequent manual filtering to remove very long sentences and sentences not representing the standard dialect of Nepali. Recordings of these excerpts, read aloud by one male and one female native speaker of Nepali, were created for use in text-to-speech applications.

The other part of the written NNC, the 'General Collection' (NNC–GS), was constructed according to rather less stringent criteria. For this dataset, we opportunistically collected as much written Nepali as possible, simply including whatever became available to us. So the NNC–GC includes the full text of numerous published books, text drawn from Nepali news websites, as well as a significant amount of data from other, printed newspapers and journals. Its final size is thirteen million words.

Due to the considerations of cost, we could not include in the NNC–GC any texts that were not already in machine-readable format. This was, in fact, one important limitation on the kinds of texts that could be included in the corpus – a limitation previously encountered by earlier projects that involved building corpora for South Asian Languages (see Baker *et al*., 2004; Hardie *et al*., 2006). It is now well-established that Unicode is the preferred choice for those encoding corpora using non-Latin alphabets (McEnery and Xiao, 2005). However, the majority of the sources of data for the NNC–GC provided text encoded in a variety of incompatible eightbit encodings (sometimes known as 'fonts'). It was therefore necessary to recode the texts as Devanagari Unicode. A methodology for this conversion process is described by Hardie (2007a), and was implemented in a series of font-converter programs by the Nelralec team of developers.

Both sections of the written corpus (NNC–CS and NNC–GC) were marked up using the XML version of the Corpus Encoding Standard[7] (XCES). It was found necessary to make some minor modifications to the XCES document type definition (DTD) to allow for all the types of structural markup which we needed to represent in the corpus; for this reason, the modified DTD is distributed with the corpus. Text metadata is stored within the XCES header of each text; metadata specific to a particular text is given in Nepali, while metadata that patterns consistently across texts is given in Nepali and English. A partial example of a text header from the NNC–CS is given in Appendix A. The main XML tags used within the body of the corpus texts are s (sentence), p (paragraph) and head (heading).

The NNC spoken corpus has been designed to follow the template of the *Göteborg Spoken Language Corpus* (see Allwood *et al*., 2003). It consists of 260,000 words of data, collected from seventeen types of social activities,

[7] See: http://www.xml-ces.org/

such as *shopping*, *discussion*, and so on.[8] We made audio-video recordings of 116 occurrences of these activities in their natural context (thirtytwo hours), and then produced annotated orthographic transcriptions (in Devanagari). However, we retain the audio-visual materials for subsequent analysis of phonetic, paralinguistic and extra-linguistic features. As well as the recordings and transcriptions, additional metadata on the recording and the participants was also collected. Along with straightforward demographic details such as sex, age and occupation, this metadata includes the native language of each speaker, their native dialect, and their second language (this last being of great importance in a community that is as multilingual as Nepal).

Finally, a substantial amount of parallel corpus data has been collected. This includes both Nepali texts with English translations, and *vice versa*. The NNC parallel corpus contains about four million words in total. This data is drawn largely from two areas: texts relating to computing and texts relating to national development issues.

The NNC has been completed only relatively recently, and we are, therefore, in the very earliest stages of exploiting the corpus for our investigation of the Nepali language. One main use of the corpus has been in lexicography. The *Samakalin Nepali Sabdakos* ('Contemporary Nepali Dictionary') has been compiled using the written part of the NNC. This, the first corpus-based dictionary of Nepali – and also of any South Asian language, to our knowledge – has initially been published online in a digital edition;[9] a subsequent, expanded version will be published in book form. The benefits of using the corpus have been immediate: for many words, new meanings have been identified which had not previously been recorded in any dictionary. While it is debatable whether a corpus of fourteen million words can be optimal for lexicography, it does appear that such a corpus is easily sufficient to make advances on non-corpus-based lexicography. One other key issue which the compilation of the dictionary has highlighted is that of spelling variation. Nepali does not yet have fully standardised spelling, and this is reflected in the corpus texts – and thus in the *Samakalin Nepali Sabdakos*.

As well as lexicography, the NNC has been exploited in NLP applications – initially, in the creation of a text-to-speech system. As noted above, the NNC speech corpus was developed specifically with this application in mind. Finally, we have begun to make use of the corpus for linguistic investigation. The NNC is used for teaching corpus linguistics within the MA degree in Linguistics at Tribhuvan University; some grammatical analysis has also been based on the

[8] The full list of activities/contexts represented in the spoken corpus is: shopping, discussion, task orientated formal meeting, task orientated informal meeting, dinner conversation, conversation while working, hotel, academic seminar, radio talk show, television talk show, interview, hospital, phone, market place, fortune telling, formal discussion and thesis defence.

[9] See: http://www.nepalisabdakos.com

corpus (see Hardie, 2007b, 2008). In order to undertake this work, software that is capable of handling the Devanagari script and the markup of the text was required. We have used Xaira[10] for this purpose; and a web-based interface to the corpus has also been developed.

3. Corpus annotation

So far, only one form of analytic annotation has been added to the corpus. That is part-of-speech (POS) tagging. In this section, we describe the process by which a POS tagset was devised for Nepali – a language for which no tagging system had previously existed – and outline the development of the automatic tagger.

3.1 Defining the tagset

The first prerequisite for part-of-speech (POS) tagging is a tagset which lists exhaustively the grammatical categories of words in the language.[11] This was developed and trialled on extracts of corpus data, undergoing a process of revision and retrialling until we were confident that the tagset could adequately cover the overwhelming majority of phenomena in unrestricted Nepali text. Many categories presented few difficulties: the tag merely needed to label a category that was already recognised in the grammatical analysis of Nepali. An example of this was the category of adverb (tagged as RR[12]); adverbs in Nepali are uninflected and this label could simply be applied to the category as traditionally analysed. However, some other aspects of the tagging presented problems which had not arisen in noncorpus-based grammatical category analysis.

One such problem was the issue of Nepali case markers. The morphemes which indicate case are typically – but not always – suffixed in the written form of the language to the noun. For this reason, we initially created a set of tags to distinguish between cases of nouns (ergative, genitive, accusative, *etc*.). However, this approach ultimately proved to be unworkable, due to certain features of the case markers. These elements may also appear attached to categories other than nouns (for instance, adjectives or verbal forms) or to one another. Multiple case markers may be found attached to a single base. Accounting for all these phenomena led to a very great proliferation in the number of tags. For the purposes of tagging, then, it was more convenient to treat the case markers as clitic postpositions. This allowed us to use a single tag for

[10] See: http://www.oucs.ox.ac.uk/rts/xaira/ and also Xiao (2006).
[11] The creation of the tagset is described in detail by Hardie *et al*. (forthcoming).
[12] The Latin-alphabet tag mnemonics are modelled on those of the C7 tagset for English (see: http://ucrel.lancs.ac.uk/claws7tags.html).

nouns (NN, or NP for proper nouns) together with tags for the various types of postposition (II, IE for the ergative marker, IA for the accusative marker, *etc*.) which were applied to the case markers as separate tokens.

The difficulties in annotating for case arose from the highly agglutinative nature of Nepali morphosyntax. This was also a factor in the tagset's analysis of the Nepali verb system. Nepali has a great range of verb inflections, each created by compounding several different verb forms, encoding a wide range of tenses, aspects and moods. Person, number and gender are also marked on verbs, as well as honorificity. As with case, tagging all of these features would have necessitated a huge number of tags (in this case, tens of thousands). This would, clearly, not be workable. Rather than retokenise the verbs, however, we adopted a scheme where the final element of any compounded verb form determined its tag. So, for instance, the verb *cha* ('is') is given the tag VVYN1 (third person non-honorific singular verb). Compounds such as *garcha* ('does, is doing') therefore also receive the VVYN1 tag. This simplified the process of annotating verbs to the point where it was a manageable task, although verb forms still account for twenty nine out of the total 112 tags. One hundred and sixty texts from the NNC–CS were annotated manually using this tagset. This data then served as the basis for the training of an automatic tagger, as described in the following section.

3.2 Automated tagging

The tagging of the NNC was accomplished using a retrained form of the *Unitag* tagging system (originally developed to tag Urdu; see Hardie, 2004, 2005). This system uses separate programs for tokenisation, analysis of tokens, and contextual disambiguation of ambiguously-tagged words. Some extensions to the system, as well as an entirely new set of linguistic knowledge resources, were required to adapt Unitag to Nepali.

As noted above, our tagging scheme involved retokenisation of postpositions separately from the nouns that they are affixed to. To accomplish this computationally, a powerful tokeniser, controlled by a list of rules in a specially-designed formalism, was created. However, no set of rules is without exceptions. For example, *k¯a*, one form of the genitive marker, is usually split apart when it is seen at the end of a word. However, in words such as *amerik¯a* ('America'), the *k¯a* should *not* be split. For this reason, the tokeniser uses a lexicon of exceptions to limit its application of the rules.

A tagger also requires lexical resources, listing the possible tags for each word in the lexicon. A lexicon for Nepali, containing around 40,000 types, was derived automatically from the manually-annotated data.

However, for some of the most common word-forms, this lexicon was less than ideal. For example, *k¯a* should nearly always be tagged IKO (genitive postposition, plural/honorific). However, its entry in the lexicon actually has many more tags. Most of these represent errors in the manually-annotated

data – in over 300,000 words, some errors are to be expected, and the high frequency of this postposition means that many of those errors will affect it. We did not wish errors such as this to be passed on in the automated tagging of the remainder of the corpus. However, we did not have sufficient time for the analyst to edit the entire lexicon manually. As an alternative to this, we manually created a smaller, second lexicon (containing approximately 300 words, and covering most of the closed-class words in Nepali) and incorporated it into the system in such a way that entries in this manual lexicon always override entries in the automatic lexicon.

For the analysis of word-forms not found by a search of the lexicon, we created a suffixlist – that is, a lexicon of word-end letter sequences, each of which implies a single tag or a small number of tags for the word on which it appears. (For example, the ending *–cha* implies the tag VVYN1, for reasons explained above.) To expedite this phase of the work, the suffixlist was generated semi-automatically from the manually annotated data, although it was necessary to postedit it by hand.

Generating the lexicon and suffixlist from corpus data allowed each tag in these databases to be allocated a probability based on its frequency in the corpus, which was a significant advantage for the implementation of probabilistic disambiguation. The main technique which we used for disambiguation was a Markov model,[13] again trained using the manually annotated data to learn bigram tag transition probabilities. The combination of the lexicons, suffixlist, and transition matrix gave an ultimate accuracy rate of around 93 percent on written Nepali data; we have not yet been able to assess the tagger's accuracy on spoken data, but expect it to be slightly lower.[14]

4. Future directions

In this final section, we will outline some of the research directions which we anticipate will prove fruitful now that we are able to exploit the NNC for the investigation of the Nepali language.

We anticipate that further applications in NLP will be possible: as well as the text-to-speech system already developed, speech-to-text and speech-to-speech systems are possible using the data we have collected. Furthermore, the parallel corpus data has clear applications in machine translation. The NNC also has many potential educational applications. The parallel corpus will be used as a basis for devising language teaching materials, and in combination with our other work on lexicography, will also assist in the preparation of future bilingual (Nepali–English) dictionaries.

In terms of the purely linguistic exploitation of the data, we intend to continue the corpus-based grammatical investigation which we have begun.

[13] For background on POS tagging using Markov models, see El-Beze and Merialdo (1999).
[14] The tagger is available for download at: http://www.lancs.ac.uk/staff/hardiea/nepali/

Key tags in fiction		*Key tags in 'general prose'*		*Key tags in 'learned'*		*Key tags in 'press'*	
More common	Less common	More common	Less common	More common	Less common	More common	Less common
PMX	JX	MM	VVYN1F	JX	PMX	NP	PMX
TT	NN	FS	TT	IKO	VVMX1	JX	VVMX1
VVMX1	NP	NN	IKF	CC	PMXKM	FB	DDX
VVYN1F	IKO	IH	VDX	FS	PTM	VE	FS
IKF	IH	JX	VVYM1F	NN	VQ	IKO	TT
VVYM1F	FS	CC	IE	IKM	VVTX2	IKM	VQ
PTM	CC	FZ	RK	FZ	VDX	NN	IFK
VDX	II	II	VE	II	IE	PXH	RR
VQ	IKM	VN	DDX	MOX	JM	IH	VVYN1F
DDX	MM	VVYX2	NP	CBS	PXH	CSA	PMXKM
RK	VN	FO	PMX	FU	MLX	IE	VVYX2
VVYN	FB	FF	PTM	MOM	VCM	VI	VVYN1
PMXKM	FZ	PMXKO	VVYN1	IKX	PTH	VN	PTH
UU	FO	VDO	IA	JT	UU	II	VVTX2
PTH	MOX	DJX	PTN		RK		JM
VVTX2	MOM	DGM	VVMX1		TT		VCM

Table 2: Key POS tags analysis across genres of the NNC–CS

However, the corpus also lends itself well to contrastive studies – both between Nepali and other languages, taking advantage of the highly comparable sampling frame of the NNC–CS, and between different types of Nepali using the various sections of the corpus. For example, we wish to undertake a corpus-based investigation of the stylistic features of different genres within the corpus, and also a detailed comparison of the written and spoken datasets. The design of the spoken corpus also enables both a comparison of language use across the different social activities represented in the corpus, and a first analysis of features of Nepali interaction such as turn-taking, feedback and gesturing.

An example of the type of genre analysis that the design of the corpus permits is the *key POS tag* analysis. Any POS tag whose frequency is (relatively) greater in a given subsection of the corpus than in the remainder of the corpus, to a degree that is highly statistically significant,[15] is regarded as a *key tag*. Key tags may reflect the relative prominence of different types of grammatical structure across a set of genres. The key tags for each broad text-type[16] within the NNC–CS, compared with the rest of the NNC–CS, are given in Table 2.

Certain trends in the data are immediately evident. The positively key tags for fiction include several different verb tags (V–), and pronoun tags (P– and D–).

[15] With a *p*-value of less than 0.01. The log likelihood statistic was used (see Dunning, 1993).
[16] The broad text-types are groupings of the fifteen genres in the sampling frame that are frequently used in contrastive studies (see, for instance, Leech, 2003: 224–5).

By contrast, adjectives (JX), nouns, (N–), numerals (M–) and adpositions (I–) – all categories associated with complex lexical noun phrases – are negatively key. So, fiction in Nepali is characterised, quantitatively, by a high frequency of verbs and pronominal noun phrases and a low frequency of lexical noun phrases. By contrast, learned prose is characterised by more or less the opposite. Most of the negative key tags are verbs and pronouns, and most of the positive tags represent adjectives, nouns, adpositions and numerals. The press category also has many key tags linked to lexical noun phrases, especially nouns and adpositions; but here there are both positive *and* negative key verb tags, and many pronoun categories are negatively key. Finally, general prose presents a wholly mixed picture. There are noun tags, adposition tags, verb tags and pronoun tags in both the positively and negatively key lists for this genre. In sum, a cline across the different genres emerges, from the verb-and-pronoun-heavy style of fiction, to the noun-phrase-heavy style of learned prose, with the general prose and press categories occupying an intermediate position.

There is no room here to extend this analysis or to place it in its proper context within the literature of similar investigations on other languages (but see Hudson, 1994; Mair *et al*., 2002; Rayson *et al*., 2002). Our work in analysing the corpus along these lines is ongoing. Furthermore, the NNC has been made freely available for non-profit research purposes,[17] and so it is our hope that other linguists and language engineers will be able to make use of this corpus for purposes that we may not even have thought of yet.

References

Allwood, J., L. Grönqvist, E. Ahlsén and M. Gunnarsson. 2003. 'Annotations and tools for an activity based spoken language corpus' in J. van Kuppevelt (ed.) Current and New Directions in Discourse and Dialogue, pp. 1–18. Dordrecht: Kluwer Academic Publishers.

Baker, P., A. Hardie, A. McEnery, R. Xiao, K. Bontcheva, H. Cunningham, R. Gaizauskas, O. Hamza, D. Maynard, V. Tablan, C. Ursu, B.D. Jayaram and M. Leisher. 2004. 'Corpus linguistics and South Asian languages: corpus creation and tool development', Literary and Linguistic Computing 19 (4), pp. 509–24.

Dunning, T. 1993. 'Accurate methods for the statistics of surprise and coincidence', Computational Linguistics 19 (1), pp. 61–74.

El-Beze, M. and B. Merialdo. 1999. 'Hidden Markov models' in H. van Halteren (ed.) Syntactic Wordclass Tagging, pp. 263–84. Dordrecht: Kluwer Academic Publishers.

[17] For details on obtaining the corpus, and information on relevant copyright issues, please contact the project team at: http://bhashasanchar.org/new/contact.php.

Hardie, A. 2004. The Computational Analysis of Morphosyntactic Categories in Urdu. Unpublished PhD thesis, University of Lancaster. Available online at: http://eprints.lancs.ac.uk/106/

Hardie, A. 2005. 'Automated part-of-speech analysis of Urdu: conceptual and technical issues' in Y. Yadava, G. Bhattarai, R.R. Lohani, B. Prasain and K. Parajuli (eds) Contemporary Issues in Nepalese Linguistics, pp. 48–72. Kathmandu: Linguistic Society of Nepal.

Hardie, A. 2007a. 'From legacy encodings to Unicode: the graphical and logical principles in the scripts of South Asia', Language Resources and Evaluation 41 (1), pp. 1–25.

Hardie, A. 2007b. 'Collocational properties of adpositions in Nepali and English' in M. Davies, P. Rayson, S. Hunston and P. Danielsson (eds) Proceedings of the Corpus Linguistics Conference, CL2007. University of Birmingham, UK. Available online at: http://www.corpus.bham.ac.uk/corplingproceedings07/

Hardie, A. 2008. 'A collocation-based approach to Nepali postpositions', Corpus Linguistics and Linguistic Theory 4 (1), pp. 19–62.

Hardie, A., P. Baker, A. McEnery and B.D. Jayaram. 2006. 'Corpus-building for South Asian languages' in A. Saxena and L. Borin (eds) Lesser-Known Languages in South Asia: Status and Policies, Case Studies and Applications of Information Technology, pp. 211–42. The Hague:Mouton de Gruyter.

Hardie, A., R.R. Lohani, B.N. Regmi and Y.P. Yadava. Forthcoming.'A morphosyntactic categorisation scheme for the automated analysis of Nepali'.

Hudson, R. 1994. 'About 37% of word-tokens are nouns', Language 70 (2), pp. 331–9.

Hundt, M., A. Sand and R. Siemund. 1998. Manual of Information to Accompany the Freiburg–LOB Corpus of British English ('FLOB').Englisches Seminar, Albert-Ludwigs-Universität Freiburg. Available online http://khnt.hit.uib.no/icame/manuals/flob/index.htm

Hundt, M., A. Sand and P. Skandera. 1999. Manual of Information to Accompany the Freiburg–Brown Corpus of American English ('Frown'). Englisches Seminar, Albert-Ludwigs-Universität Freiburg. Available online at: http://khnt.hit.uib.no/icame/manuals/frown/ index.htm

Leech, G. 2003. 'Modality on the move: the English modal auxiliaries 1961–1992' in R. Facchinetti, M. Krug and F. Palmer (eds) Modality in Contemporary English, pp. 224–40. Berlin: Mouton de Gruyter.

Mair, C., M. Hundt, G. Leech and N. Smith. 2002. 'Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged LOB and F-LOB corpora', International Journal of Corpus Linguistics 7 (2), pp. 245–64.

McEnery, A. and Z. Xiao. 2004. 'The Lancaster Corpus of Mandarin Chinese: a corpus for monolingual and contrastive language study', Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC) 2004, pp. 1175–8. 24–30 May 2004. Lisbon, Spain.

McEnery, T. and Z. Xiao. 2005. 'Character encoding in corpus construction' in M. Wynne (ed.) Developing Linguistic Corpora: A Guide to Good Practice. AHDS Literature, Languages and Linguistics, Oxbow Books. Rayson, P., A.Wilson and G. Leech. 2002. 'Grammatical word class variation within the British National Corpus sampler' in P. Peters, P. Collins and A. Smith (eds) New Frontiers of Corpus Research: Papers from the Twenty First International Conference on English Language Research on Computerized Corpora, Sydney 2000, pp. 295–306. Amsterdam: Rodopi.

Shastri, S.V. 1986. Manual of Information to Accompany the Kolhapur Corpus of Indian English, for use with digital computers. Available online at: http://khnt.hit.uib.no/icame/manuals/kolhapur/index.htm Xiao, Z. 2006. 'Xaira: an XML-aware indexing and retrieval architecture' (review), Corpora 1 (1), pp. 99–103.

Appendix A: Extract showing bilingual metadata in a text header from the Core Sample of the Nepali National Corpus

```
<cesHeader version="2.1">
<fileDesc>
    <titleStmt>
        <h.title>NNC-CS: sample A16</h.title>
        <respStmt>
            <respType>Electronic version created by
            </respType>
            <respName>
                Nelralec / Bhasha Sanchar Project (भाषा सञ्चार)
            </respName>
        </respStmt>
        <respStmt>
            <respType>transcribed by</respType>
            <respName>सरिता दहाल तिम्सिना</respName>
        </respStmt>
    </titleStmt>
    <extent>
        <wordCount>2005</wordCount>
        <byteCount units="kb">137</byteCount>
    </extent>
[...]
    <sourceDesc>
        <biblStruct>
            <monogr>
                <h.title>जनमत अर्द्धसाप्ताहिक</h.title>
                <imprint>
                    <publisher>सुशिला चुके</publisher>
                    <pubDate>
                        २०४८-०४-२०; २०४८-०६-१४;
                        २०४८-०८-२६; २०४८-०९-२९
                    </pubDate>
                    <pubPlace>नेपालगन्ज</pubPlace>
                </imprint>
            </monogr>
        </biblStruct>
    </sourceDesc>
</fileDesc>
[...]
```