The Kiranti comparable corpus: a prototype corpus for the comparison of Kiranti languages and mythology

Aimée Lahaussois

paru dans : Jones, Mari ed., 2014. *Endangered Languages and New Technologies*. Cambridge: Cambridge University Press, pp. 17-34

The Kiranti comparable corpus: a prototype corpus for the comparison of Kiranti languages and mythology

Aimée Lahaussois (CNRS, UMR 7597, HTL, Univ Paris Diderot, Sorbonne Paris Cité)

1. Introduction

This chapter describes the concepts and methodologies which form the basis for a prototype corpus developed with data from Khaling, Thulung and Koyi, three endangered languages of the Kiranti group of Tibeto-Burman languages, spoken in eastern Nepal. The corpus aligns versions of the same story in these three languages, tagging narrative material of similar semantic content so that it can be called up for comparison. The interface enables the data within the corpus to be viewed in several different ways, thus making it possible to compare the different lexical items and morphosyntax used in each linguistic version of the story.

The prototype corpus includes material from a single story, but will be expanded over the next few years to include many more elements from the Kiranti mythological cycles, and data from additional speakers, and eventually, it is hoped, from other Kiranti languages. The concepts and methods of parallel and comparable corpora, which until now have been limited to well-described languages, have been exploited here to carry out comparative analysis of closely related endangered and under-described languages, based on culturally authentic narrative material. This approach can be used for any language group which shares a common narrative tradition.

The corpus describe in this chapter was developed in collaboration with Séverine Guillaume, who built the technical framework for the aligned corpus (Lahaussois & Guillaume 2012). This work is part of the larger *HimalCo* project ('Parallel corpora in Himalayan Languages) funded by the French *Agence Nationale de Recherche* (2013-2015), which will involve the documentation of languages of the Naish, Rgyalrongic and Kiranti subgroups of Tibeto-Burman. The project's outcomes include the compilation of comparable corpora based on collected narrative data which will be used for linguistic comparison within and between the three subgroups. It should be stressed that what is advocated here is not a particular software configuration but, rather, a concept, the technical implementation of which could be realised in a number of different ways. This chapter aims to highlight the way that aligning comparable corpora of endangered language materials can reveal features (both narrative and morphosyntactic) that may not previously have been hitherto documented for a given variety or varieties.

The fact that the Kiranti languages share a mythological cycle is well-known to researchers working on these cultures and languages, and mythological texts are included in most descriptive grammars of the subgroup. The anthropologist N.J. Allen, who wrote a grammar of the Thulung language (Allen 1975), has written widely about Thulung mythology, placing it in a larger comparative context and tracing certain elements to pre-Buddhist Tibet and further afield (see for example Allen 1980, 1997). Allen's work on comparative mythology remains anthropological and, as such, he does not use his material to engage in any form of linguistic comparison.

In her book *The Structure of Kiranti Languages* (1994), Ebert provides a comparison of the phonology and morphosyntax of six Kiranti languages, basing her analysis on existing grammars of these languages and the texts provided in the grammars. She states that her comparative work was 'originally planned as an introduction to a volume of mythological

texts' (1994: 10) which was eventually published separately (Ebert & Gaenszle 2008). Despite the project's links with Kiranti mythology, the linguistic analysis is based on the mostly non-mythological narrative materials that are reproduced in an appendix to the work (1994: 154-280), and the linguistic comparison does not engage with the concept of shared narrative tradition.

In a subsequent work, *Camling Texts and Glossary* (2000), Ebert presents, *inter alia*, three versions of the Khocilipa story in Camling. She lays out the main narrative events of the story, relates which parts are reproduced in which dialect, compares these with versions of the story available in different Kiranti languages, and presents the interlinearized and translated Camling texts. Her work, appears to be the first to compare compare different versions of the same Kiranti mythological text (as opposed to Allen, who compares their themes and features). Ebert's alignment data are presented by listing the correspondences between sentences that occur in the three Camling versions of the story (2000: 8). Ebert did not have access to tools that would have allowed her to align the texts digitally and her work differs from the corpus presented here in that her main interest seems to have resided in comparing the narrative structure of the different versions of each story rather than in using the aligned material for purposes of comparative linguistic analysis.

Ebert and Gaenszle (2008) revisits the body of shared Kiranti mythology, taking into account all the languages for which mythological narrative data has been collected. Building on previous work, Gaenszle, provides an analysis of the common structure and content of the four major cycles, namely myths of creation, myths about the culture hero, myths of ancestral migration, and myths about first settlements and village foundations. Ebert's contribution (2008: 17-50) on the grammars of the languages chosen for analysis does not differ substantially from her earlier (1994) work. Although many of the illustrations are drawn from the mythological cycle, the individual examples do not match up in terms of narrative event.

The fact that the material is drawn from a shared mythology is not relevant to the way in which it is used for linguistic comparison. For example, although the sentences chosen to illustrate topic marking are both from they same story (in different languages) (Ebert 2008: 37), they come from very different parts of the story and are, consequently, no more useful for the comparison of shared linguistic features than if they had no narrative relationship whatsoever.

The Kiranti comparable corpus described here represents a significant departure from previous work on Kiranti languages. This is due, in large part, to the improvement in corpus tools that new technologies can offer. The project involves building a digital corpus, which can be analyzed using corpus tools, such as a concordancer. It contains data of similar narrative content which, although mainly mythological in nature, are generally no more closely related than collections of stories from different traditions would be. Data within the corpus are aligned: this matches up lexical and morphosyntactic similarities between languages and allows all versions of the story to be viewed together. By providing a large corpus made up of multiple parallel stories in different languages it is hoped that the project will make it possible to establish facts about different narrative traditions within the different Kiranti subgroup and to develop a better sense of how the linguistic features of these endangered languages compare with each other.

2. The Kiranti languages

The Kiranti subgroup of Tibeto-Burman languages is comprised of some thirty languages spoken in eastern Nepal by small groups of several thousand speakers (see Figure 1). All but one, Libu, have an exclusively oral tradition. A number of these languages have been the subject of descriptive grammars, all but one of which have been written within the framework

of the Himalayan Languages Project (http://www.himalayanlanguages.org/): Wambule (Opgenort 2004), Jero (Opgenort 2005), Kulung (Tolsma 2006), Sunwar (Borchers 2008), and Bantawa (Doornenbal 2009).[i] Other projects on this language group, such as the Chintang language research program (http://www.spw.uzh.ch/clrp/), also promise to increase our knowledge of the Kiranti languages and our access to spontaneous narrative material.

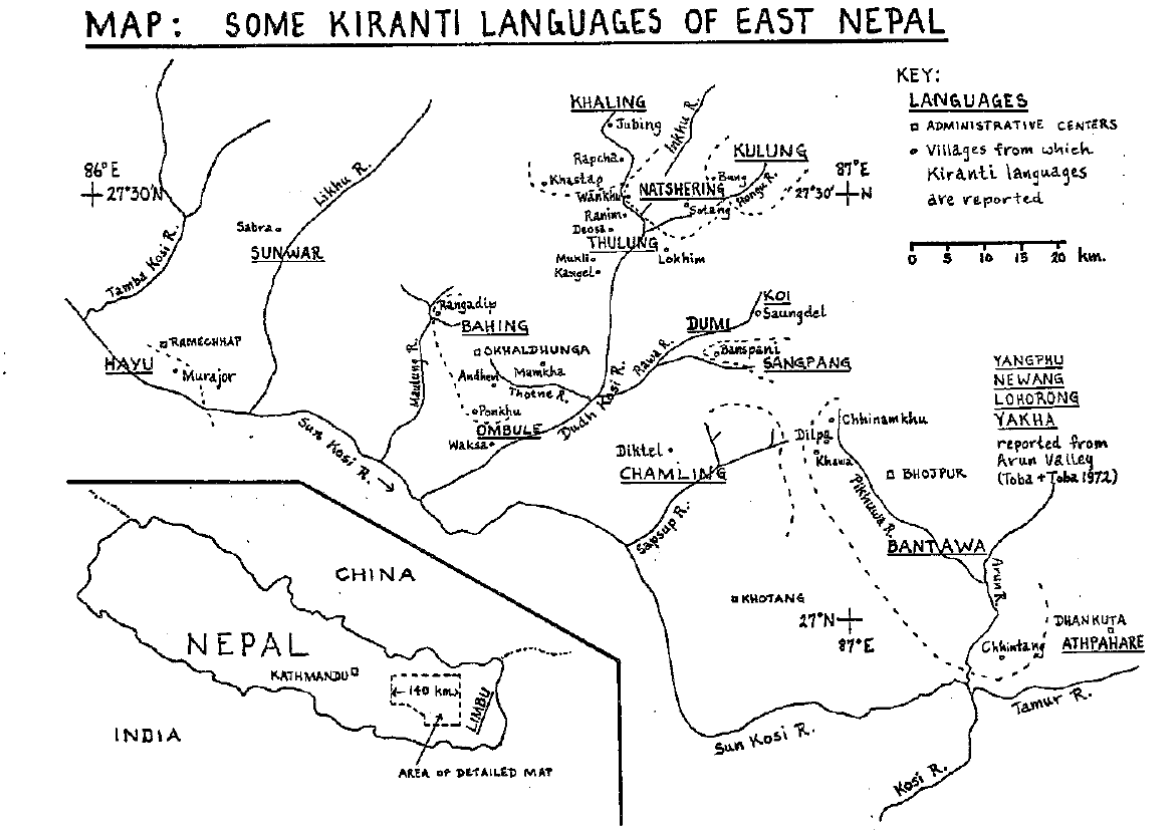

Figure 1. Map of the Kiranti area (after Michailovsky 1975)

It is not clear to what degree the languages in this family are related. Michailovsky has published phonological reconstructions of initial consonants for proto-Kiranti (2009), which strongly suggests that they represent a genetic subgroup. However, Ebert claims that 'It has never been shown that Kiranti [..] is a valid genetic unit. [...] Hansson assumes in an unpublished report of the Survey Project [Linguistic Survey of Nepal] that the cluster of

Kiranti languages results from several migration waves of Tibeto-Burman groups that have influenced each other for a longer period'. (Ebert, 2003: 516). While the prototype corpus presented here is too small to help provide answers to questions of this sort, the enhanced Kiranti comparable corpus, once enriched with additional stories, speakers and languages, may well provide tools which will make it possible to gain a better sense of how closely these different languages are related.

3. Parallel vs. comparable corpora

In the field of translation studies, translational corpora are aligned in such a way that translation equivalents can not only be viewed and compared easily, but also recalled in order to facilitate future translation tasks. This method of aligning linguistic material has been adopted by a number of typologists who need a tool that can enable them to compare the features of different languages (cf. Cysouw & Wälchli 2007). Examples of large translation-based corpora include works such as *Le Petit Prince*, the Harry Potter series, the Bible and European parliamentary texts. The materials are aligned with software which, using punctuation and multilingual dictionaries, proposes automatic alignments which are then corrected by users. Despite the fact that these are translated versions of the same text, sometimes difficulties in aligning the material can still occur. For example, Stolz (2007: 105) notes that 'For the translations of *Le Petit Prince* [...], identical length can only be achieved by cutting off the text at a pre-determined mark because the languages differ widely as to the number of pages, words, or sentences they use'.

Despite the difficulties in aligning even translational equivalents, the term 'parallel corpus' is widely used to describe such materials. Sinclair (1996) proposes the following practical definition: 'A parallel corpus is a collection of texts, each of which is translated into

one or more other languages than the original'. Wälchli (2007: 132) identifies the numerous biases which users of parallel corpora must take into account: '(a) written language bias [...], (b) bias toward planned (conscious) language use (including purism) [...], (c) bias toward religious and legalese registers, (d) narrative register bias, (e) bias toward large languages (in spread zones), (f) bias toward standardized (simplified?) language varieties, (g) bias toward non-native use of languages, (h) bias toward translated language (rather than original language use).'

In an attempt to correct for these, comparable corpora have also been developed. A comparable corpus is defined as one which 'selects similar texts in more than one language or variety, [with] as yet no agreement on the nature of the similarity. [...] The possibilities of a comparable corpus are to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus' (Sinclair 1996). An example of texts that constitute a comparable corpus might be different language versions of news reports about a same political or sporting event, where the content is similar but, as a result of being produced directly in the target language, the texts are not distorted by translation. In languages with an established written tradition, comparable corpora can build up a large volume of similar texts, which are then automatically aligned using algorithms. Parallel corpora depend by definition on the existence of translational materials and are therefore inevitably more limited in volume.

For the Kiranti languages, the shared mythological cycle (similar, native versions of stories, and, crucially, not translation-derived) appears to favour the compilation of a comparable corpus. It does, however, differ from traditional comparable corpora in terms of its small volume of data. Moreover, as the Kiranti languages do not have a written tradition, the tools which are typically used for automatic alignment (electronic dictionaries, parsers) are not available.

The popularity of using stimulus material such as *Frog, where are you* (Meyer 1969) and the *Pear Story* (see Chafe 1980), to collect typologica data means that, for these stories, data are available for a large number of languages. Although these are good materials to use for purposes of comparative linguistic study in the sense that the data collected are produced by native speaker and do not suffer from any translation-related biases, they are not, arguably, truly authentic since they result from a visual input that can be interpreted differently from one person to the next. Such a situation also holds true for speakers of unwritten languages, for whom the interpretation of printed or video images may be so unfamiliar as to lead to rather unusual narratives. This is pointed out by Stolz and Stolz (2008: 33): 'Recording free discourse and/or narrations of picture-book stories may lead to multilingual corpora which are too diverse both structurally and semantically to allow for direct comparison because one cannot be sure that the data at hand are compatible with one another'.

The Kiranti comparable corpus presents a good solution to the problems discussed above. Synchronically-speaking, it is not translation-derived and it is truly 'authentic', in that the stories are culturally and linguistically autochthonous rather than derived from picture-books or videos.[ii] The corpus is thus lexically, morphosyntactically and pragmatically representative of the Kiranti languages, and well-suited to linguistic analysis with the aim to revealing characteristic features and constructions of the languages under study.

4. Source data for the Kiranti comparable corpus prototype

In order to establish the prototype for the comparable corpus, a story which had been collected in three different Kiranti languages was chosen, namely that of Kakcilip (the Thulung name for the main character). Gaenszle calls this the 'culture hero' cycle (1991: 248, 2008: 6) and provides a description of the main narrative elements, based on the Mewahang

version of the story (1991:271-288) and on other Kiranti versions to which he has had access (2008: 8-9). The story may be summarized as follows:

- The hero is a descendant of the First Man;

- He is always depicted as an orphan living with his two sisters;

- The sisters and brother separate, after the brother appears to have died;

- The boy survives through cunning;

- When fishing, he catches a stone repeatedly, which turns out to be a woman, who becomes his wife;

- After building a house, the brother summons his sisters with the help of various animals.

The prototype corpus is made up of a Thulung, Khaling and Koyi version of this story.[iii] The Thulung and Khaling stories are of roughly equivalent length (transcribed audio recordings of twelve and thirteen minutes respectively), while the Koyi version is considerably longer (sixty-three minutes) because it was narrated as part of a complete foundation myth. In the interest of preserving the integrity of the original source materials, it was decided to use the Koyi narrative in its entirety, aligning only the pieces that correspond to the Kakcilip story with the material from the other languages.

The data in the corpus is interlinearized using *Interlinear Text Editor*, a software package developed by the LACITO research group in order to generate an appropriate format for archiving in the Pangloss Collection (formerly the LACITO Archive). The data consist, classically, of a transcription tier, a glossing tier, and a translation tier, along with audio tags that synchronize sound data with each sentence unit. As the data that make up the Thulung and Koyi versions of the story were already archived, it was decided that, in building the corpus, the original source files should not be modified. As a result, information about the alignment between the different stories making up the corpus is encoded in an additional

document, ('alignment file'), which establishes links between different sentences in each story but, crucially, without affecting the original source files.

The comparable corpus is thus made up of two different types of file:

a) *annotation files*, of which there is one per language version of each story. The information contained in these files includes the transcription, glossing and translation of the individual sentence, word, and morpheme units that make up the text (see Jacobson et al. 2001; Thieberger & Jacobson 2010).

b) *alignment files*, of which there is one per story, identifying the links that exist between elements contained in the different language versions. Alignment files are created using a spreadsheet: the different versions of the story are manually lined up in pairs, and the corresponding sentences are identified and labeled as similar. This information is then converted into xml in order to generate the alignment file (Lahaussois &Guillaume 2012: 34). The alignment phase needs the notion of correspondence between sentences to be defined. This is discuss below.

5. The notion of comparability in the corpus

The alignment of the corpus is based on the concept that certain segments can be compared to others, and that this revolves around the notion of similarity. Note that in defining comparable corpora, Sinclair (1996) points out that there is 'as yet no agreement on the nature of the similarity'. In the case of the Kiranti comparable corpus, a similarity is defined as a Segment, represented by one or more sentences that contain material of similar narrative function or content.

As a result of such a definition, we can establish a typology of similarities found in the corpus, based on whether the similarity is one of function or content, namely:

a) similarities with shared narrative function only

b) similarities with shared narrative content

c) similarities with shared morphosyntactic constructions

5.1. Similarities with shared narrative function only

These similarities link sequences within the narrative which serve the same narrative purpose, even though, linguistically, they may share nothing else. This may be illustrated by (1), which represents an important turning point in the narrative of the Culture hero story (see Section 4), namely the passage where the sisters and brother separate after the brother appears to have died. This episode is related in the Thulung and Khaling versions of the story, but rather differently in each case: in one language, the two sisters believe their brother (who is asleep) to be dead and build a bamboo hut to cover his remains, while in the other, the sisters inadvertently bury their sleeping brother with nettle peelings while they are working and, when they cannot find him, they assume that he has died. The episode has shared narrative function, as it represents the starting point for separate brother-and-sister-adventures. However, the content is not shared. The differences in content are even more striking when examined in detail (for all examples, the language is identified with a three-letter code (THU for Thulung, KHA for Khaling and KOY for Koyi); gloss abbreviations are provided at the end of the chapter):

(1)

THU

əni	meɖɖa-m	pəʈsʰi	kolem	ʈsʰipʤi-kam	nem	bɤne-saka
and	then-NMLZ	after	one.day	cut.bamboo-GEN	house	make-CVB

mɯ-gunu	u-ri	kʰaktsilip-lai	am-saka	
that-inside	3SG.POSS-sibling	Kakcilip-DAT	make.sleep-CVB	

‘Then they made a house out of pieces of big bamboo, and put their brother Kakcilip to sleep inside it’.

KHA

grômmɛ-kolo	lasmɛ-su-ʔɛ	dhawa	mɛ	ʤʌkhʌl	kâ:k-tɛsu-	
Gromme-COM	Lasme-DU-ERG	quickly	that	nettle.fibre	peel-3DU>3SG.PST-	
lo	mɛ	lektsêm-ʔɛ	nek-to	nek-to	khɵs-tɛ	
TEMP	that	nettle.core-INS	cover-CVB	cover-CVB	go-3SG.PST	

‘Gromme and Lasme quickly peeled the nettle fibre and covered him with the inside of the fibre’.

Note that, in the Thulung version of the story, the sisters are referred to with a pronoun (the possessive prefix in *u-ri*, ‘their brother’) and the brother, by name. In the Khaling version of the story, the sisters are referred to by name, and the brother by a demonstrative (*mɛ*, ‘that’). Moreover, the material with which the brother is covered is different: bamboo in the Thulung version, and nettle fibre in the Khaling version. The differences between this pair of sentences are such that they pose a serious problem for automatic alignment, as there are no similar lexical elements. Nevertheless, it seems important to align these segments in order to be able to use the corpus for research of a wider scope. Once the corpus is enlarged beyond the current prototype, it is possible that other versions of the story (by different speakers, in

different languages or dialects) will reveal that the similarity highlighted above does in fact share more elements than those that are currently available. In other words, in the absence of shared linguistic material among the different versions of the story, the segments should still be aligned: first, because the alignment will ultimately be expanded to include other languages and different versions of the same text within the same language which may involve elements that help bridge the differences we see here; and second, because of the possibility that the corpus may be used by non-linguists, who need alignment of more than lexical correspondence (for example, anthropologists may be interested in looking for potential ethnographically-relevant differences).

5.2. Similarities with shared narrative content

In this type of similarity, the sentences not only refer to the same event within the narrative, but also express that event with shared lexical items. The linguistic similarities are mostly lexical, but there are sometimes also grammatical morphemes which are cognate or functionally similar.

The sentences in (2), for example, relate the same event in the story, namely the fact that the protagonists becoming orphans. These sentences share lexical items, such as 'orphan', 'become', 'be', and also a few grammatical elements, such as the intransitive 3PL.PST agreement marker, and clause-combining morphology, such as the sequential marker *-ma* in Thulung and temporal marker *-lo* in Khaling which, though different phonetically, are still relevant for the comparison of how such markers combine with finite verb forms and sequence clauses.

(2)

THU

mɯrmim-kam	tin	ʤana	ba-mri	ʦɤŋɖatura	
3PL-GEN	three	person	be-3PL.PST	later	orphan
dym-miri-ma	ba-mri				
become-3PL.PST-SEQ	be-3PL.PST				

‘The three of them were there and later became orphans’.

KHA

grômmɛ	lasmɛ	khakʦalʌp	ʦəttsə	mō:-tnu-lo
Gromme	Lasme	Kakcalop	children	be-3PL.PST-TEMP
reskʌp	ʦhʉk-tɛnu			
orphan	become-3PL.PST			

‘When Gromme, Lasme and Kakcalop were children, they became orphans’.

This type of similarity is useful for comparing lexical items within the languages, and their specific usages in context. Such information is made even easier to retrieve when it is accessed using the concordancer (see section 6.2). These similarities also give us information about basic sentence construction.

5.3. Similarities with shared morphosyntactic constructions

Where sentences are identified as sharing a construction, the alignment reveals the morphosyntactic features of the languages being compared. This is exemplified in (3) and (4):

(3) imperative form for 2SG agent with 1SG patient, coupled with a direct speech construction

THU

ɖiʈ-ŋi	by-ry
leave-2SG>1SG.NPST	do-3SG>3SG.PST

‘ “Leave me”, she said’.

KOY

leʔ-ʦu	dja
leave-2SG>1SG.IMP	say.3SG>3SG.PST

‘ “Leave me”, she said’.

(4) complement clause construction, involving the same lexical material

KOY

nana-nusi-ja	mind-usi	ʦʰa	ɔ-bɔkʦi	miʦ-a
o.sister-DU-ERG	think-3DU.PST HS		1SG.POSS-y.sibling	die-3SG.PST

‘The sisters thought: “Our brother has died”’.

KHA

mʌnʌ khaktʂalʌp mis-tɛ mimsî-iti

then Kakcilip die-3SG.PST think-3DU.PST

‘Then they thought: “Kakcilip has died”’.

Comparing such constructions in this way can, of course, also reveal information about the grammar of the languages under examination.

This three-part typology of similarities gives a sense of the range of comparable material that exists within the corpus, as well as of what is meant by ‘similarity’ in this context. Although the notion of ‘similarity’ inevitably contains some subjectivity, once the corpus is sufficiently built up, and includes with several versions of each story in every language variety, it will provide an important source of comparative material on the languages in question.

6. Tools for viewing and analysing the corpus

This section presents the different tools which are built into the corpus interface and which allow data to be retrieved for purposes of comparison and analysis.

6.1. Viewing

The corpus interface is designed to allow two ways of viewing material. The first, called the Integral Text View, is the basic view that is seen when the corpus is opened. In the Integral Text View, each version of a story appears in its integral form in a column. For the prototype,

this means that the full Thulung, Khaling, and Koyi versions of the story appear in columns side by side (see Figure 2).

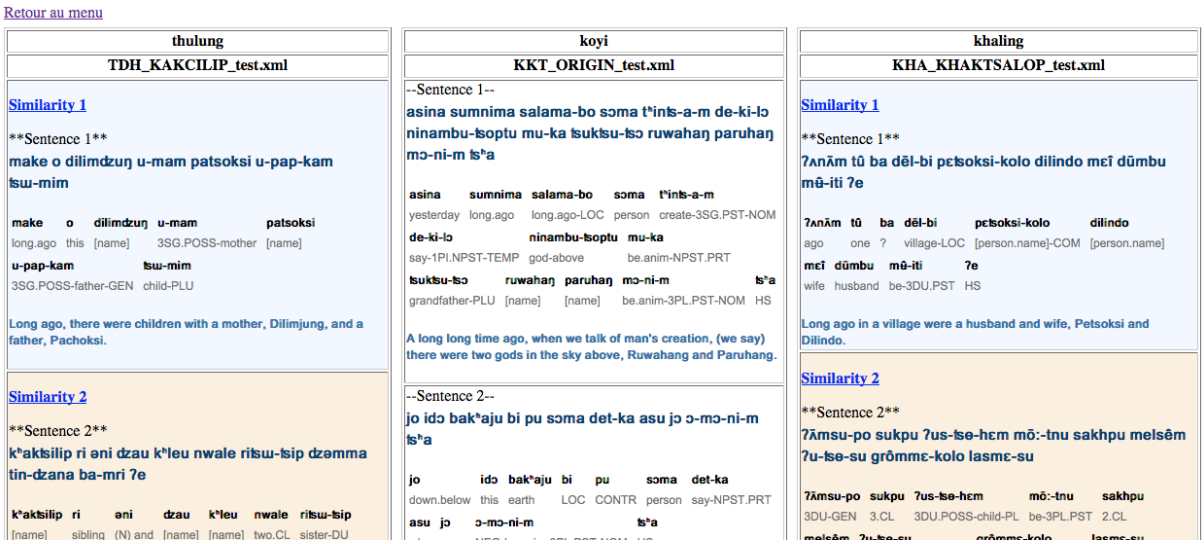

Figure 2. The Integral Text View

The idea behind the Integral Text View is that users are able to read the entire text of one language version of a story by scanning down the column. Although a certain proportion of the material in any given story will not have equivalents in the others, and will therefore not be aligned in terms of their similarities, the data are presented nonetheless, in order to maintain the narrative and morphosyntactic integrity of each version of the story. Where similarities exist between the different language versions, these are signalled by a hyperlinked label ('Similarity #') and are identified by colour so that, when scrolling through the text, one can identify visually which sentences participate in a given similarity and what those correspondences are. The colour identification is important since the order of the similar segments may differ from one version of the story to another.

The second way of viewing is called the Similarity View, which is displayed when one of the similarity labels are selected in any of the stories: it shows the equivalent sentence or sentences in the different language versions of that story. In some cases, only two languages are involved in a similarity, while in other cases, the similarity involves all three languages.

Figure 3. The Similarity View (for reasons of space, the Khaling version of the similarity is omitted here)

The Similarity View is where the real analysis of differences between the languages becomes possible. The sentences identified as sharing a similarity are aligned in such a way as to allow a deeper glimpse of how the different languages in the sample express similar narrative content.

One important issue that raises its head at this juncture is the necessity for consistent morphosyntactic glossing across the versions in order to identify with relative ease how each language expresses a particular construction. For the prototype corpus, a single field researcher was involved, thus reducing the potential differences in glossing between the

versions. As stories (and eventually other languages) are added to the corpus, this consistency will need to be maintained. Implementing glossing standards, such as the Leipzig Glossing Rules (http://www.eva.mpg.de/lingua/resources/glossing-rules.php), can help with this.

6.2. The Concordancer

A concordancer is built into the corpus interface. It can be used to perform searches on either the glossing tier, by looking up any English word or morphological gloss, or the transcription tier, by looking up a specific morpheme in any one of the languages.

The results are generated as a table, such as that exemplified in Figure 4. The left and right contexts for the term are given and also identification codes for the language, story, sentence number of each occurrence. Clicking on the highlighted term under *Mot* ('word') opens the Similarity View for that sentence and its equivalents in the other languages.

Retour au menu

Rechercher un terme :

OK

Texte	Phrase	Contexte gauche	Mot	Contexte droit	Gloses
TDH_KAKCILIP_test.xml	s15	kərəŋ ŋa ʂɤk ʂi sat kərəŋ ʂɤk ʂi	si	ra lɤ ʂi rak ʂi m guʂi huʈ	die
TDH_KAKCILIP_test.xml	s17	lɤk ʂi m pəʂʰi memlo kʰakʂilip ʂʰəhi me	si	saka hunuktʰjo ələkgai ba saka rep tʰaɖ ɖy	die
TDH_KAKCILIP_test.xml	s86	mɯʈʂip ka ʂʰəhi mɯ ʣau kʰleu ka ʂʰəhi	si	p ljak pa ri ku nɤŋ sɯ pa	die
KKT_ORIGIN_test.xml	s132	kubimɔpa ne dʰaiʔwɔ dja di aŋ heɲe m	miʔ	mu bʰak tɔ m bela heɲe m jok	die
KKT_ORIGIN_test.xml	s204	nana nusi ja mind usi ʂʰa ɔ bɔkʂi	miʂ	a kim nɔ jo tʰɔm bi kʰuʂ asi	die
KKT_ORIGIN_test.xml	s220	nɔ sɔ pu buwa leʔ si ʂʰa an	miʂ	ena an ja ebo ʂʰoʔ na miʂ ena	die
KKT_ORIGIN_test.xml	s220	an miʂ ena an ja ebo ʂʰoʔ na	miʂ	ena kʰo idɔ a buwa ŋal e aŋ	die
KKT_ORIGIN_test.xml	s242	kʰoʔ nɔ da ʂʰa da m me pu	miʂ	a ŋi dʰanɔ dʰam lɔ kim kʰoʔ nɔ	die
KKT_ORIGIN_test.xml	s329	sɔmɔ kʰɔmʂ asina habo ŋoʔ mu ma mama	miʂ	a kʰojɔ habo ŋoʔ mu ma re ki	die
KKT_ORIGIN_test.xml	s351	ʂʰaŋgara pʰiŋ usi ʂʰa ʂʰaŋgara jɔ ja tʰja	miʂ	a ʂʰa dʰaiʔ mu mu to ne pɔma	die
KKT_ORIGIN_test.xml	s364	ja ne dja si ʂʰa inʂi bɔkʂi ne	miʂ	a dja si kʰa lukʰe dja si m	die
KKT_ORIGIN_test.xml	s364	dja si kʰa lukʰe dja si m a	miʔ	e mo ʂʰa lɔ sɔmɔ kʰoʂ isi nɔ	die
KHA_KHAKTSALOP_test.xml	s29	mu thʉk wɛ mu ʣɛ̈: wɛ mʌnʌ khakʂalʌp	mis	tɛ mimsî iti mɛjûŋ lekʂêm ʔɛ nek tʉ	die

Figure 4. Concordance results for the English term 'to die'.

The concordancer enables equivalent English translations or morphological glosses to be added to different languages and also (using IPA transcription) for any phoneme or

sequence of phonemes to be searched for It can additionally be used to generate multilingual glossaries, which provide not only the equivalent lexical items across the Kiranti languages in the corpus, but also example sentences to illustrate each of the terms. Furthermore, as the audio files are synchronized with the transcription, the multilingual glossaries can form the basis of 'talking dictionaries', with sound clips provided to illustrate the pronunciation of each entry and example sentence.

7. Some results

The small size of the prototype corpus limits the amount of comparison that can currently be carried out. However, promising signs have emerged of what will possible once the corpus has been enlarged. This section discusses two results which give a sense of the type of analysis the corpus makes possible.

7.1. The identification of language-internal variation

In order to explore how comitative marking interacts with dual marking, a concordance of the gloss 'COM' was performed on the corpus. The Similarity View of the search results revealed the alignment of sentences given in (5):

(5)

KOY

runtsʰis-wa	dʰep-nasi-nɔ	mɔ	tsʰa	sul-
winnowing.basket-INS	cover-3SG.PST.REFL-SEQ	be.anim.3SG.PST	HS	hide-
nasi	tsʰa			

3SG.PST.REFL HS

‘He covered himself with a basket and stayed there and hid’.

THU

naŋlo-num	kutso-num	ʤer-tʰɑk-y	kʰrems-ɖa	ba-
winnowing.basket-COM	broom-COM	hold-hide-3SG>3SG.PST	cover-3SG.PST	be-

iɖa-m

3SG.PST-NMLZ

‘He held and hid with the basket and broom and covered himself’.

Both these sentences relate the same episode within the story, and the winnowing basket appears as an instrument in both. However, in the Koyi sentence, the instrumental marker is used, whereas the Thulung sentence indicates the instrument by the comitative marker. This is surprising, as the comitative marks accompaniment more generically via an animate object, rather than an instrumental. In other words, this similarity pairing enabled revealed that the comitative marker can also be used with inanimate objects. In other words, using the comparable corpus made it possible to identify language internal variation, through comparison with other languages in this under-documented, endangered language.

7.2. The identification of potential errors of analysis

The comparable corpus also makes it possible to identify potential errors of analysis. The sentences in (6) and (7) both refer to the moment in the narrative when the hero, weak from hunger and thirst, falls asleep, leading to his sisters' assumption that he is dead.

In the Khaling version of the story, both 'hunger' and 'thirst' were marked for the instrumental, and this is reflecgted in the glosses.

(6)

KHA

sôː-ʔɛ mʌt-tɛ-na kʉmîn-ʔɛ mʌt-tɛ-na

hunger-INS have.to-3SG.PST-SEQ thirst-INS have.to-3SG.PST-SEQ

ʔip-dək-tɛ-m

sleep-AUX-3SG.PST-NMLZ

'He was hungry and thirsty and had fallen asleep'.

However, when the Khaling version of the sentence is compared with the Koyi equivalent, it becomes clear that the Koyi term for 'hunger' was transcribed and glossed as a single lexical item, without instrumental marking. And yet, the word ends in a syllable identical to the Koyi instrumental marker, which is *-wa*.

(7)

KOY

ʤimu a-dʰoʔd-u ne soʔwa dʰal-ʤa soʔwa

food	NEG-find-3SG>3SG.PST	TOP	hunger	sway-DUR.3SG.PST	hunger
dʰal-dza-lɔ	ne	ipʰ-a-suʈs-a		ʈsʰa	
sway-DUR.3SG.PST-TEMP	TOP	sleep-COPY-AUX-3SG.PST		HS	

'When he could not find food, he swayed from hunger, when he swayed from hunger, he fell asleep'.

It is possible that the word was not properly analysed, and that it is indeed made up of the lexeme 'hunger' plus the instrumental marker. It goes without saying that this needs to be rechecked in the field, but whether or not it turns out to be an analysis error, this finding highlights another of the corpus' strengths, namely as an additional tool for checking transcription and analysis through comparison with closely related languages.

8. Conclusion

The next phase of the project will be to add more stories to the corpus. The longer-term goal is to add other Kiranti languages to the corpus. The HimalCo project will apply the methodology described in this chapter to the Rgyalrongic and Naish languages of China. Alignment will be used to study the following three areas:

a) intra-speaker variation (single speaker, different versions of a narrative)

Alexis Michaud working on Naish languages spoken in China, plans to use the alignment to compare several versions of the same story recounted by a single speaker. Data for this type of work is produced when ofr example, speakers recording a version of a story in a given linguistic variety, suddenly claim that it is 'no good' and ask to 'try again'.

b) inter-speaker variation (same dialect/language, different speakers).

This is similar to what was attempted by Ebert (2000) for several versions of the same story in Camling, and can help determine how the presence or absence of narrative elements differ when the same story is told by different speakers. For example, does the narrative structure differ according to dialect group or language? Are these differences related to the geographical distribution of the elements within the story, or are they the result of idiosyncracries in speakers' personal versions of the story?

c) inter-language variation (different languages within the same subgroup and across subgroups)

Ultimately, our goal is to compare the insights derived from the use of a comparable corpus for a given linguistic subgroup across many different subgroups.

Its planned future development will allow the corpus to view similarities according to a number of different criteria. As the main menu will list the different stories available and the versions recorded by different speakers, in different dialects and languages, users will be able to specify the criteria that interest them and build a sub-corpus which reflects those interests. The alignment files will ensure that the resulting sub-corpus retains all the information about similarities across its constituent material. The ability to build a personalized sub-corpus has the potential to provide many new insights about the connections that may exist between the Kiranti languages.

In considering the medium-term future of the methodology presented in this contribution, the following trends seem relevant:

a) Shifting from language description to language documentation

With the shift of emphasis from language description to language documentation that has occurred over the last decade, the trend seems to be towards collecting and presenting data

with the aim of making these widely accessible, both in terms of their physical availability (such as the development of open-access online archives) and in their re-use for interdisciplinary purposes. In France, this trend is reflected in the development of funding programs and of the research infrastructure: for example, the French *Agence Nationale de Recherche* has a funding scheme that is specifically directed at supporting the development of projects within the digital humanities, such as the compilation of multi-use corpora and tools. Structural initiatives that favour work on corpora include the *Written Corpora* consortium (itself part of the Corpus infrastructure, http://www.corpus-ir.fr/). This body is organized into working groups, one of which is specifically aimed at bringing together researchers involved in the compilation of multilingual corpora.

b) Developing tools for under-resourced languages

The biennial Language Resources and Evaluation Conferences (LREC) are a good predictor of current research in computational linguistics that is being applied within the broader discipline. Increasing numbers of workshops at these conferences point to a growing interest in under-resourced and endangered languages.

Moreover, institutional efforts are also underway to ensure that all languages are better represented in cyberspace: UNESCO's Communication and Information sector is tasked with, *inter alia*, facilitating Internet access and the development of digital tools for less widely known languages. This may mean that the technical difficulties linguists currently face when building corpora (aligning data manually, creating an interface, chosing appropriate data formats) will be resolved as increasing numbers of tools are developed, leaving linguists to concentrate on the actual data.

c) Accessing linguistic data:

As more and more Kiranti languages are described and analysed, the availability of data on this family is likely to increase over the next two decades. This should result in increasingly larger data samples, which will enhance the corpus, extend it to other languages and narratives, and make it even more useful for comparative study. The current emphasis on a digital format for data in linguistic projects means that narrative corpora and digital dictionaries of these languages will probably be developed as part of future documentation projects. These digital materials will make it easier to automatize the alignment of the corpus, and to increase its size.

The three trends discussed above suggest that the Kiranti comparable corpus is likely to enjoy a certain longevity. Crucially, it allows access to rare data in a novel way. Gaenszle (2008: 11) has pointed out the gaps in our knowledge about Kiranti mythology: 'Given that we lack a large corpus of myths told by different persons, it is difficult to see whether the lack of an episode in one telling is a feature of the local tradition or simply the result of the narrator's mood that day' (my translation). Once enhanced, as planned, to include additional languages and multiple speakers and dialects for each language, the Kiranti comparable in the above statement. Morevoer, it is hoped that the methodology developed for the Kiranti comparable corpus will ultimately be applied to other language groups that share a narrative tradition.

Gloss abbreviations

AUX, auxiliary; COM, comitative; CVB, converb; DAT, dative; DU, dual; DUR, durative; ERG, ergative; GEN, genitive; HS, hearsay; IMP, imperative; INS, instrumental; NEG, negative; NMLZ, nominalizer; NPST, non-past; PL, plural; POSS, possessive; PST, past; REFL, reflexive; SEQ, sequencer; SG, singular; TEMP, temporal; TOP, topic; X>Y, agent X acting on patient Y

[i] Doornebal (2009) was not written as part of the Himalayan Languages project.

[ii] It is of course possible that, in the past, stories may have been borrowed from one language into the other.

[iii] I would like to acknowledge the funding awarded for my fieldwork from the Fulbright Foundation, the Hans Rausing Endangered Language Documentation Programme, and the LACITO research group.