A viewing and processing tool for the analysis of a comparable corpus of Kiranti mythology

Aimée Lahaussois, Séverine Guillaume

A viewing and processing tool for the analysis of a comparable corpus of Kiranti mythology

Aimée Lahaussois*, Séverine Guillaume**
*CNRS, UMR 7597, HTL, Univ Paris Diderot, Sorbonne Paris Cité, F-75013 Paris
**CNRS, UMR 7107, LACITO
E-mail: aimee.lahaussois@linguist.jussieu.fr, guillaume@vjf.cnrs.fr

Abstract

This presentation describes a trilingual corpus of three endangered languages of the Kiranti group (Tibeto-Burman family) from Eastern Nepal. The languages, which are exclusively oral, share a rich mythology, and it is thus possible to build a corpus of the same native narrative material in the three languages. The segments of similar semantic content are tagged with a "similarity" label to identify correspondences among the three language versions of the story. An interface has been developed to allow these similarities to be viewed together, in order to allow make possible comparison of the different lexical and morphosyntactic features of each language. A concordancer makes it possible to see the various occurrences of words or glosses, and to further compare and contrast the languages.

Keywords: trilingual comparable corpus, Kiranti languages, mythological narrative cycle

1. Introduction

The challenges encountered when using various stimulus materials to generate parallel or similar texts for language comparison are well-known: "Recording free discourse and/or narrations of picture-book stories may lead to multi-lingual 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." (Stolz and Stolz, 2008: 33). In reaction to this, we became interested in the idea of using *native* stories in different languages as the basis for comparative work. Languages of the Kiranti group of Eastern Nepal share a very rich mythology (Ebert and Gaenszle, 2009) which can be used for this purpose. The stories are remarkably similar, both in their content and, in some cases, in their use of idiosyncratic morphosyntax which is otherwise difficult to elicit.

The Kiranti languages of Eastern Nepal are in the Tibeto-Burman family. There are two major subdivisions within the group: Limbu, on the one hand, is the language in the group with the largest number of speakers and a writing system; the 30-odd Rai languages make up the rest of the group. The Rai languages are exclusively oral[1], and spoken by small communities usually numbering several thousand speakers. They are severely endangered, due to the inroads of the national language Nepali.

While there have been a number of descriptions of Rai languages[2], there has been very little comparative work, except on a case by case basis. Ebert (1994, 2003) has written about the shared structure of the Kiranti languages, and Michailovsky (2009) has carried out work on the phonological reconstruction of proto-Kiranti, but on the whole, comparative work is limited, both in number of languages (a sample of six for Ebert's 1994 comparative work) and also in scope.

The body of shared mythology among Rai peoples presents itself as an appealing option for carrying out comparison work. The ubiquity of the mythological cycle as a form of narrative becomes apparent quite quickly to anyone working on the documentation of these languages. Most spontaneously told stories will be drawn from this body, and the stories are remarkably similar across languages.

Our goal in this paper is to describe how we have created a prototype for a Kiranti comparable corpus by aligning the same story, taken from the mythological cycle, in three languages from the group in order to advance and enable comparative work among these languages.

2. The Kiranti comparable corpus

The data presented in this paper is from personal fieldwork on three languages of the Kiranti subgroup, namely Thulung, Koyi and Khaling[3]. The creation of a comparable corpus could also be achieved using materials in existing descriptions of other Kiranti languages. Such grammars contain transcribed oral narrative which almost invariably includes elements of the same mythological cycle.

For our prototype comparable corpus, we chose to use a single story, with the goal of building up the corpus to include new stories as we collect and align them. The basic storyline for the story which we selected is the following:

[1] Any references to Rai "texts" in this paper are to transcriptions (by the linguistic researcher) of oral narratives.

[2] One can cite for example the grammars that have come out of the Himalayan Languages Project (Sunwar, Wambule, Jero, Yamphu, Bantawa, Dumi,...).

[3] The Khaling data comes from fieldwork done in collaboration with Dhana Bahadur Khaling, Guillaume Jacques, Boyd Michailovsky, Martine Mazaudon, Marie-Caroline Pons.

> Kakcilip and his two olders sisters are orphaned and must learn to fend for themselves. Kakcilip, being the youngest, is not able to contribute much, and his sisters take on the bulk of the work. One day, while they are out in the forest, Kakcilip falls asleep. The sisters, thinking he is dead, leave him behind and decide to separate, each flying in a different direction. One of the sisters encounters an owl, who eats her. The other sister comes looking for her, and manages to get her bones back from the owl. With an enchantment, she rewakens her sister and explains what has happened. Later on, the sisters encounter a series of animals--a louse, a flea, a goat and finally a cock which calls out "khakcilipa" when it comes near them. They realize this is a sign from their brother, as the cock is calling his name, and follow him back to a place where they are reunited. Kakcilip has in the meantime had an adventure of his own in which he, while fishing, caught a stone which turned out to be a female figure he eventually marries.

In some cases, the story was narrated as an independent story: this is the case in Thulung and Khaling. In Koyi, it is woven into a very long origin myth. Thus our stories are all of different lengths[4] (Thulung: 12 minutes, Khaling: 13 minutes, Koyi: 63 minutes), with the Koyi version contained a large amount of additional material which is not in the other two stories.

3. Building the corpus

The corpus was built from preexisting interlinearized XML "annotation" files of the Kakcilip story in three languages. These files were in a format which is used by the LACITO Archive (http://lacito.vjf.cnrs.fr/archivage/index_en.htm) and contain three tiers of data (transcription into IPA, glosses, and free translation), as is typical of analyzed field data used in the description of oral and endangered languages. In the case of all three languages in our corpus, this three-tiered structure was generated using interlinearization software called ITE (Interlinear Text Editor) developed specifically for the LACITO Archive by Michel Jacobson.

Because each language's XML annotation files for the Kakcilip story are archived, we decided, in compiling our corpus, to preserve the original format of the files, rather than modify them to include alignment data. We therefore decided to create a distinct alignment file, in which we defined similar segments, which we call "similarities", across the different texts making up the corpus. A similarity is defined here as a segment, represented by one or more sentences, containing material of similar narrative content or function. Our definition is thus based on narrative and not lexical or morphosyntactic criteria. While we would have prefered a configuration where the basis for similarity alignments was more linguistically-oriented criteria, this was not possible considering the spontaneously produced narrative data we had to work with. Unsurprisingly though, passages of similar narrative content often contain lexical material and structures that are close and sometimes even identical, so that in effect our narratively-based alignment proves useful for linguistic analysis.

The similarities were identified manually by reading through each of the texts in language pairs (Thulung-Koyi, Koyi-Khaling, Thulung-Khaling) and recording into a spreadsheet which sentence numbers of each text corresponded, in semantic content, to which others, and assigning to each correspondence a similarity label.

The spreadsheet was then converted into XML using a perl script, as illustrated in Figure 1.[5]

The annotation files called up by the alignment file contain information about the content of each of viewing levels (users can chose to look at the data in Text, Word or Morpheme views) generated by the ITE software. The text (<TEXT>) breaks down into sentences (<S>) which in turn break down into words (<W>) and morphemes (<M>). Each unit can contain a transcription (<FORM>) and a translation or gloss (<TRANSL>. This is illustrated in Figure 2.

The comparable corpus is thus made up of four files: the three languages' annotation files, which contain the entire version of the story in each language, and an alignment file in XML which contains the information laying out the correspondences between the language versions.

We then defined a graphic interface making it possible to view the alignments of sentences. Considering that a priority for endangered language documentation is often the widespread diffusion of data, we decided to use web-related technology. PHP and XSL style sheets were created to view the corpus.

The first viewing option is of the individual texts in their entirety, with one language per column. We call this the "integral text view". Similarities are identified by a color scheme, so that they can be identified across languages at a glance. This was important because, owing to the great differences in length between the Koyi version of the story and the other two language versions, and the different ordering of narrative events, the similarities rarely occur on the same page in all three languages. The "integral text view" is illustrated in Figure 3.

The second viewing option allows the user to select one of the similarities, and see all the content which corresponds to it in the different languages. We call this the "similarity view", and it is obtained by clicking on any similarity label in any of the three stories. The similarity view is illustrated in Figure 4.

Each of these viewing options has a related XSL style sheet and uses a PHP program to switch from one view to the other.

[4] The standard for comparison is the durations of the language versions, as the transcriptions which make up the comparable corpus are of oral recordings.

[5] Figure 1 and all subsequent figures are found at the end of the article.

We have also developed a concordancer which makes it possible to search for any word or gloss found within the transcription and glossing tiers respectively. Figure 5 shows the results of a concordance on the gloss "sister". The results show the transcription tier, with the word corresponding to the concordanced gloss highlighted (regardless of whether the search was for a word or a gloss). The sentence and source text for each result are identified (the "text" label in the left-most column identifies the story, starting with its Ethnologue language code, TDH=Thulung, KKT=Koyi, KHA=Khaling), and the left and right context for the term are given. The concordance function in effect generates a trilingual correspondence for any gloss in the corpus, and is a useful way to build up a trilingual glossary. This function will be more useful as the corpus is expanded to include more stories covering a greater narrative (and therefore lexical) range.

Each occurrence can be selected (by clicking on the highlighted word) and opens the similarity view: the sentence, if it is part of a similarity set, is shown together with the corresponding sentences in other languages. This makes it possible to identify the morphosyntactic constructions used to expressing the same narrative content.

A concordance of the gloss "INS" (instrumental marker) leads, among other results, to Similarity 35 (which, in the interest of space, is reproduced not as a screen shot but as the text which makes up the similarity, namely examples (1) and (2) below):

(1) [THU][6]
naŋlo-**nuŋ** kutso-nuŋ
winnowing.basket-COM broom-COM
ʤer-tʰɑk-y kʰrems-ɖa
hold-hide-3SG>3SG.PST cover-3SG.PST
ba-iɖa-m
be-3SG.PST-NMLZ
'He held and hid with the basket and broom and covered himself.'

(2) [KOY]
runtsʰis-**wa** dʰep-nasi-nɔ
winnowing.basket-INS cover-3SG.PST.REFL-SEQ
mɔ tsʰa sul-nasi tsʰa
be.anim.3SG.PST HS hide-3SG.PST.REFL HS
'He covered himself with a basket and stayed there and hid.'

Where Koyi uses an instrumental marker (*-wa*) to encode the semantic role of the instrument (the winnowing basket Kakcilip is using to hide himself), Thulung unexpectedly uses a comitative marker (instead of instrumental marker *-ka*), usually reserved to express accompaniment by a person. This type of example points to the potential usefulness of this corpus in uncovering, through comparison, language-internal variation which would not necessarily be covered in descriptive grammars.

4. Issues encountered

4.1 Methodological issues

A number of issues were encountered during the construction of the comparable corpus, including methodological questions about the necessity for manual alignment of the texts, and the nature of similarities. These are discussed below.

4.1.1 Hand alignment

The identification and definition of similarities in the material must be carried out manually. From our understanding, the tools available for well-described languages with numerous digital resources (dictionaries, POS taggers, etc) cannot be used to automatize the work we have done with the Kiranti corpus. This is precisely one of the significant differences between so-called mainstream languages and little described minority ones. The matter of hand-alignment does not represent a problem in the case of the Kiranti corpus, as we are dealing with very small data sets. Nonetheless it will be necessary as the corpus grows to include other languages to find methods to partially automatize the alignment.

4.1.2. The typology of similarity judgments

As defined in section 3 above, similarity judgments were based on the degree of narrative similarity of textual segments, and were thus inherently subjective. Because the three versions of the story are close, and because of the proximity of these languages, similarities often involve equivalent lexical items and sometimes even the same morphosyntactic constructions, but not always. Some examples will be given of the three basic types of similarities we have found.

Similarities with only narrative function in common

Similarity 5 aligns sentences which share almost nothing but narrative function. There is not a single word which is the same across the languages, and grammatically, the only shared element is the use of a converbal marker (*-saka* in Thulung, *-to* in Khaling), as seen in examples (3) and (4) below.

(3) [THU]
əni meɖɖa-m pətsʰi kolem tsʰipʤi-kam nem
and then-NMLZ after one.day cut.bamboo-GEN house
bɤne-saka mɯ-gunu u-ri kʰaktsilip-lai
make-CVB that-inside 3SG.POSS-sibling Kakcilip-DAT
am-saka
make.sleep-CVB
'Then after they made a house out of pieces of big bamboo, and put their brother Kakcilip to sleep inside it.'

(4) [KHA]

[6] All examples will be preceded by a three-letter abbreviation of the language name: THU for Thulung, KOY for Koyi and KHA for Khaling.

grômmɛ-kolo lasmɛ-su-ʔɛ dhawa mɛ ʤʌkhʌl
Gromme-COM Lasme-DU-ERG quickly that nettle.fiber
kâ:k-tɛsu-lo mɛ lektsêm-ʔɛ
peel-3DU>3SG.PST-TEMP that nettle.core-INS
nek-to nek-to khɵs-tɛ
cover-CVB cover-CVB go-3SG.PST
'Gromme and Lasme quickly peeled the nettle fiber and covered him with the inside of the fiber.'

The bamboo in one version of the story is nettle fiber in the other; Kakcilip is mentioned by name in one verions but not the other; the house which covers Kakcilip in one version is a pile of fiber in the other. And yet, narratively, this is the point at which their brother gets covered--because they think he is dead in the Thulung version, and because he is sleeping and they do not see him in the Khaling version--and at which the sisters and brother begin to live their separate stories. Linguistically, this similarity brings us very little, but it could be useful for, for example, an ethnographic study of the evolution, across Kiranti tribes, of basic household activities (the story makes clear that the bamboo- and nettle-peeling are a fundamental household chore).

Similarities with narrative content and some lexical material in common

Similarity 3 aligns sentences which share narrative content (it refers to the point at which the protagonists become orphans) and also some lexical and grammatical material, as seen in examples (5) and (6).

(5) [THU]
mɯrmim-kam tin ʤana ba-mri tsɤŋɖa tura
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.'

(6) [KHA]
grômmɛ lasmɛ khaktsalʌp tsɵttsɵ mō:-tnu-lo
Gromme Lasme Kakcalop children be-3PL.PST-TEMP
reskʌp tshɯk-tɛnu
orphan become-3PL.PST
'The children Gromme, Lasme and Kakcalop were there and became orphans.'

These two sentences contain examples of existential predication; both use clause sequencing morphosyntax (*-ma* for Thulung, *-lo* for Khaling) and they share lexical items "orphan" and "become" (the latter with a 3rd plural past conjugation in both languages). Again, this is not earth-shattering, linguistically, but provides interesting information.

Similarities revealing shared grammatical constructions

In other cases, the alignments turn up some shared linguistic constructions.

Similarity 4 (examples (7) and (8)) reveals an identical construction for "to come to a decision, to advise with each other", which we find in both Khaling and Thulung here. In Thulung, it involves a loan word from Nepali (*salla*) but in both cases it involves the verb "to do", and we see that in both languages, the agents are ergative-marked. This is a construction that does not come up naturally in elicitation, and the fact that it emerges from the data suggests that there is something to be gained from an alignment based on narrative content.

(7) [THU]
utsi-walwak-ka ʤau-nuŋ kʰleu-nuŋ-ka
3DU.POSS-sibling-ERG Jau-COM Khleu-COM-ERG
tsʰəhi səlla bet-tsi ʔe
CONTR advice do-3DU>3SG.PST HS
'Jau and Khleu came to a decision.'

(8) [KHA]
tunɵl didi bahini grômmɛ
one.day older.sister younger.sister Gromme
lasmɛ-su-ʔɛ mɵl mɯ-ssu
Lasme-DU-ERG counsel do-3DU>3.PST
'One day, Gromme and Lasme had a discussion.'

Similarity 7 (examples (9) and (10)) brings up two elements of interest: the lexical items "hunger" and also the construction "to fall asleep" which, in both languages, contains an additional aspect-bearing element (the auxiliary verbs *sutʂ-* in KOY and *dɵk-* in KHA) which, again, does not come up unless in an appropriate context. An additional element of interest here is that *soʔwa* (in example (9)), elicited in Koyi as a single word, appears to be a mistake: looking at the Khaling cognate and at how the word is used in Khaling suggests that the Koyi equivalent should probably have been analyzed as *soʔ-wa* (hunger-INS). This remains to be verified with a native speaker, but would point to a potential additional benefit of the multilingual alignment if it helps refine transcription and analysis.

(9) [KOY]
ʤimu a-dʰoʔd-u ne soʔwa
food NEG-find-3SG>3SG.PST TOP hunger
dʰal-ʤa soʔwa dʰal-ʤa-lɔ
sway-DUR.3SG.PST hunger sway-DUR.3SG.PST-TEMP
ne ipʰ-a-sutʂ-a tsʰa
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.'

(10) [KHA]
sô:-ʔɛ mʌt-tɛ-na kɯmîn-ʔɛ
hunger-INS have.to-3SG.PST-SEQ thirst-INS
mʌt-tɛ-na ʔip-dɵk-tɛ-m
have.to-3SG.PST-SEQ sleep-AUX-3SG.PST-NMLZ
'He was hungry and thirsty and had fallen asleep.'

4.1.3. Minor issues

A number of other minor issues were identified, which are part and parcel of the alignment of any material across languages.

-It is important that the glosses used across the languages of the corpus be consistent, in order to simplify concordancing. Even though the three versions of the story were analyzed and glossed by the same person, there are some inconsistencies that must be corrected.

-The similar content for one segment is only found in two of the languages and not the third: this was of course a minor problem, and inevitable given the different narrative structures of the three versions of the story. The alignment file records sentence number information as long as at least two languages share any one similarity.

-The chosen unit for identification of similarities is the sentence, yet only part of the sentence contains similar material across languages. Some similarities thus look like they contain very different material. It was nonetheless felt to be important that any similarities be identified, even if they only involved a small part of a sentence, as any similarity could be relevant for linguistic comparison.

-The order in which the similarities occur within each narrative is different across languages. We resolved this issue by using different colors for each similarity, in order to be able to identify them visually at a glance, and by making it possible to call up a specific similarity's content in the three languages by clicking on the similarity label. (The result is what we see in Figure 4).

4.2. Comparable vs parallel corpus?

One interesting consideration is whether we are dealing with materials for a comparable or a parallel corpus in this instance. If we take the basic definitions laid out in the EAGLES report on corpora, "a parallel corpus is a collection of texts, each of which is translated into one or more other languages than the original." This definition is opposed to that for a comparable corpus, "which selects similar texts in more than one language or variety, [with] as yet no agreement on the nature of the similarity." (Sinclair, 1996). On the one hand, the texts are not identical, something demonstrated very quickly when trying to align the segments. This would suggest that these materials make up a comparable corpus. As a general rule, though, languages have quite different ways of encoding information, resulting in different lengths for a same text, suggesting that no two texts, even when they result from translation, can ever be truly parallel. Note Stolz's (2007: 105) comments about the Petit Prince multilingual corpus: "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." One of the main issues in determining whether we must consider this a comparable or a parallel corpus is that the bulk of theoretical work on corpora seems to involve written materials. In the case of oral materials, which can contain all manner of production errors and self-corrections, it is difficult to imagine that two narratives could ever be "parallel", even if they are by the same speaker. And yet the material, at a metalinguistic level, is the same. To cite Maia (2003) "comparability is in the eye of the beholder."

One of the reasons the questions is even relevant is that there is some debate about whether the Kiranti languages constitute a genetic grouping or instead a cluster of languages that have been in contact for a very long time within a cultural area. "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). Is the material making up our corpus an ancestral proto-Kiranti mythological cycle which has been transmitted through time into successive generations of daughter languages (in which case it is originally the *same text*) or have these stories been transmitted through cultural borrowing among languages which look close but are perhaps not genetically related (despite what looks like a fair amount of shared vocabulary--see Michailovsky (2009) for proto-Kiranti reconstructions), in which case our different language versions of the story constitute *translations* of the original? These questions of genetic grouping and inheritance may well be what this corpus enables to get closer to resolving: lexically, the languages look quite close, but structurally, much more analysis is needed.

5. Avenues opened by such a comparable corpus

The methodology proposed in this paper should in principle be applicable to other languages and subgroups, as long as narratives can be found which are common to the languages to be compared. The main goal, as we conceive it, is essentially linguistic: we aim to find narrative materials that can reveal significant aspects of the (morpho)syntax of the language studied in its own terms.

One such project is currently underway using the Kiranti comparable corpus: a study of the scope of dual and comitative marking, of their combination with other case markers, and coocurrence with numerals and classifiers. The corpus seems well adapted to such a study, and the data so far gives evidence of considerable variation. One appealing aspect of the multilingual corpus is that the similarities reveal unexpected uses, such as seen in examples (1) and (2), where a concordance for comitative markers revealed the use of an instrumental marker in one of the languages.

The Kiranti corpus fits into a larger project, in collaboration with Guillaume Jacques and Alexis Michaud, of building comparable corpora for three subgroups of Tibeto-Burman languages from the greater Himalayan region (Kiranti, rGyalrong and Na). While only Kiranti languages have shared native mythology, the rGyalrong and Na languages have folklore (borrowed from Tibetan in the case of rGyalrongic languages) which

would provide rich materials for the building of such a comparable corpus.

One other angle that we would like to explore is the extension of this concept of comparable corpus to different configurations[7]:

1) multiple versions of a same story by a single speaker (intra-speaker variation)

2) multiple speakers of a same dialect

3) multiple speakers of different dialects of the same language

In addition to the possibilities the comparable corpus opens for linguistic analysis and comparison, there is a strong potential for use by ethnographers documenting oral reports of different customs across a number of communities.

6. Conclusion

While work on endangered languages has embraced the possibilities of corpus linguistics for some time, we feel that our multilingual comparable corpus, which has the crucial distinction of being built of native narrative materials, represents a new tool in the arsenal of the linguist wishing to do comparative work on underdescribed languages.

The size of the comparable corpus presented here is very small (as is natural considering the labor-intensive nature of data collection, transcription, glossing, translation and sound-synchronization, usually involving a single linguist), but will be expanded with additional matching texts and additional languages in the group. This type of comparable corpus will make a larger-scale comparison of the Kiranti languages, which has been limited, more feasible.

The small size of the corpus, the necessity of manual alignment (of a sometimes subjective nature), may be countered by the fact that it does not suffer from most of the biases of larger parallel corpora of more mainstream languages. Wälchli (2007: 133) cites the following biases for the use of parallel text corpora for typological research: "(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)."

The only one of these biases which can be leveled against the Kiranti comparable corpus is (d), namely "narrative register bias", as all the material is from a single narrative register. The Kiranti corpus is based exclusively on transcribed oral narrative material; it is made up of foundational mythological texts which cannot be claimed to be religious (or legal) in nature. The languages are spoken by at most several thousand people in a mountainous region, and the corpus is thus made up of truly "minority" material for which there is no standardized language variety (standardization seems to be the domain of written languages, and endangered languages show "an additional layer of variation" even among oral tradition languages (Grinevald, 2007: 45)). As the corpus does not involve translation (the free translation in the data is associated to each sentence by the linguist after data collection) and therefore represents native language use.[8]

The avoidance of so many of the biases against parallel corpora is very strongly in the favor of a comparable corpus such as we have produced. There seems to be enough evidence of the potential usefulness of the corpus and viewing and analysis tool that we feel it to be worthwhile to continue to build the corpus, initially with additional texts already in our possession, and later on by including data from other languages. We feel that the Kiranti comparable corpus may ultimately provide a means of getting a better sense of the linguistic variation (both internal and cross-linguistic) in Kiranti languages, and perhaps offer evidence towards deciding whether or not this is genetic grouping.

7. References

Davies, A., (2003). *The native speaker: myth and reality*. Clevedon: Multilingual Matters

Ebert, K., (1994). *The Structure of the Kiranti Languages*. Arbeiten des Seminar fur Allgemeine Sprachwissenschaft Nr. 13. Zürich: Universitat Zürich Seminar fur Allgemeine Sprachwissenschaft.

Ebert, K., (2003). Kiranti languages: an overview. In: Thurgood G. and R. LaPolla (eds.), *The Sino-Tibetan Languages*. London and New York: Routledge, pp. 505-517.

Ebert, K. and Gaenszle, M., (2009). *Rai mythology: Kiranti Oral Texts* (Harvard Oriental Series, 69). Cambridge: Harvard university press.

Grinevald, C., (2007). Encounters at the brink: linguistic fieldwork among speakers of endangered languages. In: Miyaoka, O., Sakiyama, O. & Krauss, M. (eds), *The Vanishing Languages of the Pacific Rim*. Oxford, Oxford University Press

Jacobson, M., Interlinear Text Editor: http://michel.jacobson.free.fr/ITE/index_en.html

Lewis, P.M., (ed.), (2009). *Ethnologue: Languages of the World*, Sixteenth edition. Dallas, Tex.: SIL International. Online version**:** http://www.ethnologue.com/**.**

Maia, B., (2003). What are comparable corpora? Electronic resource, found at

[7] This idea comes from Guillaume Jacques and Alexis Michaud (pc)

[8] The question of who deserves to be called a "native" speaker is a fairly complex issue (Davies 2003), all the more so in situations of extreme endangerement and intense contact.

http://web.letras.up.pt/bhsmaia/belinda/pubs/CL2003%20workshop.doc

Michailovsky, B., (1975). Notes on the Kiranti Verb <East Nepal>. *Linguistics of the Tibeto-Burman Area* 2.2. pp.183-218.

Michailovsky, B., (2009). Preliminaries to the comparative study of the Kiranti subgroup of Tibeto-Burman. *Proceedings of the International Symposium on Sino-Tibetan Comparative Studies in the 21st Century, June 24-25, 2010*. Academia Sinica, Taipei, Taiwan. pp. 145-70.

Sinclair, J., (1996). Preliminary recommendations on Corpus Typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards).

Stolz, T., (2007). Harry Potter meets Le Petit Prince: On the usefulness of parallel corpora in cross-linguistic investigations. In: Cysouw, M & Wälchli, B. (eds.), Parallel Texts: Using Translational Equivalents in Linguistic Typology. Theme issue of Sprachtypologie und Universalienforschung STUF 60.2: 100-117.

Stolz, C. and Stolz, T., (2008). Functional-typological Approaches to Parallel and Comparable Corpora: the Bremen Mixed Corpus. LREC 2008: Workshop on building and using comparable corpora. 33-38.

Cysouw, M. and Wälchli, B. (2007). Parallel texts: using translational equivalents in linguistic typology. In: Cysouw, M. & Wälchli, B. (eds.), Parallel Texts: Using Translational Equivalents in Linguistic Typology. Theme issue of Sprachtypologie und Universalienforschung STUF 60.2: 95-99.

Wälchli, B., (2007). Advantages and disadvantages of using parallel texts in typological investigations. In: Cysouw, M. & Wälchli, B. (eds.), Parallel Texts: Using Translational Equivalents in Linguistic Typology. Theme issue of Sprachtypologie und Universalienforschung STUF 60.2: 118-134.

8. Acknowledments, abbreviations

We would like to thank various institutions and agencies which provided funding for field research on the three languages in our corpus: the Fulbright Foundation, the Hans Rausing Endangered Language Documentation Program, and LACITO-CNRS.

Abbreviations: The Leipzig Glossing Rules have been applied, along with including additional abbreviations where needed:
AUX, auxiliary; COM, comitative; CONTR, contrastive; CVB, converb; DAT, dative; DU, dual; DUR, durative; ERG, ergative; GEN, genitive; HS, hearsay; INS, instrumental; NEG, negative; NPST, non-past; PL, plural; POSS, possessive; PST, past; REFL, reflexive; SEQ, sequencer; SG, singular; TEMP, temporal; TOP, topic; X>Y, indicates agent X acting on patient Y

```
<similarities>
	<files>
		<file xml="TDH_KAKCILIP_test.xml" lang="thulung" sound="../audio/Kakcilip.wav"/>
		<file xml="KKT_ORIGIN_test.xml" lang="koyi" sound="../audio/Origin.wav"/>
		<file xml="KHA_KHAKTSALOP_test.xml" lang="khaling" sound="../audio/Khaktsalop.wav"/>
	</files>
	<similarity id="1">
		<color>aliceblue</color>
		<file id="TDH_KAKCILIP_test.xml">
			<sentence id="s1"/>
		</file>
		<file id="KHA_KHAKTSALOP_test.xml">
			<sentence id="s1"/>
		</file>
	</similarity>
	<similarity id="2">
		<color>antiquewhite</color>
		<file id="TDH_KAKCILIP_test.xml">
			<sentence id="s2"/>
		</file>
		<file id="KKT_ORIGIN_test.xml">
			<sentence id="s191"/>
		</file>
		<file id="KHA_KHAKTSALOP_test.xml">
			<sentence id="s2"/>
			<sentence id="s3"/>
			<sentence id="s4"/>
		</file>
	</similarity>
</similarities>
```

Figure 1. Alignment file, generated from a similarity alignment spreadshee

```
<TEXT  xml:lang="x-sil-tdh" id="crdo-TDH_KAKCILIP">
<S id="s1">
   <AUDIO start="1.1704" end="12.0457"/>
     <FORM kindOf="phono">make o dilimʤuŋ u-mam patsoksi u-pap-kam tsɯ-mim</FORM>
   <TRANSL xml:lang="en">Long ago, there were children with a mother, Dilimjung, and a father,
Pachoksi.</TRANSL>
   <W><M><FORM kindOf="phono">make</FORM>
     <TRANSL xml:lang="en">long.ago</TRANSL>
     </M>
   </W>
   <W><M><FORM kindOf="phono">o</FORM>
     <TRANSL xml:lang="en">this</TRANSL>
     </M>
   </W>
   <W><M><FORM kindOf="phono">dilimʤuŋ</FORM>
     <TRANSL xml:lang="en">[name]</TRANSL>
     </M>
   </W> ……..
```

Figure 2. Contents of part of an annotation file

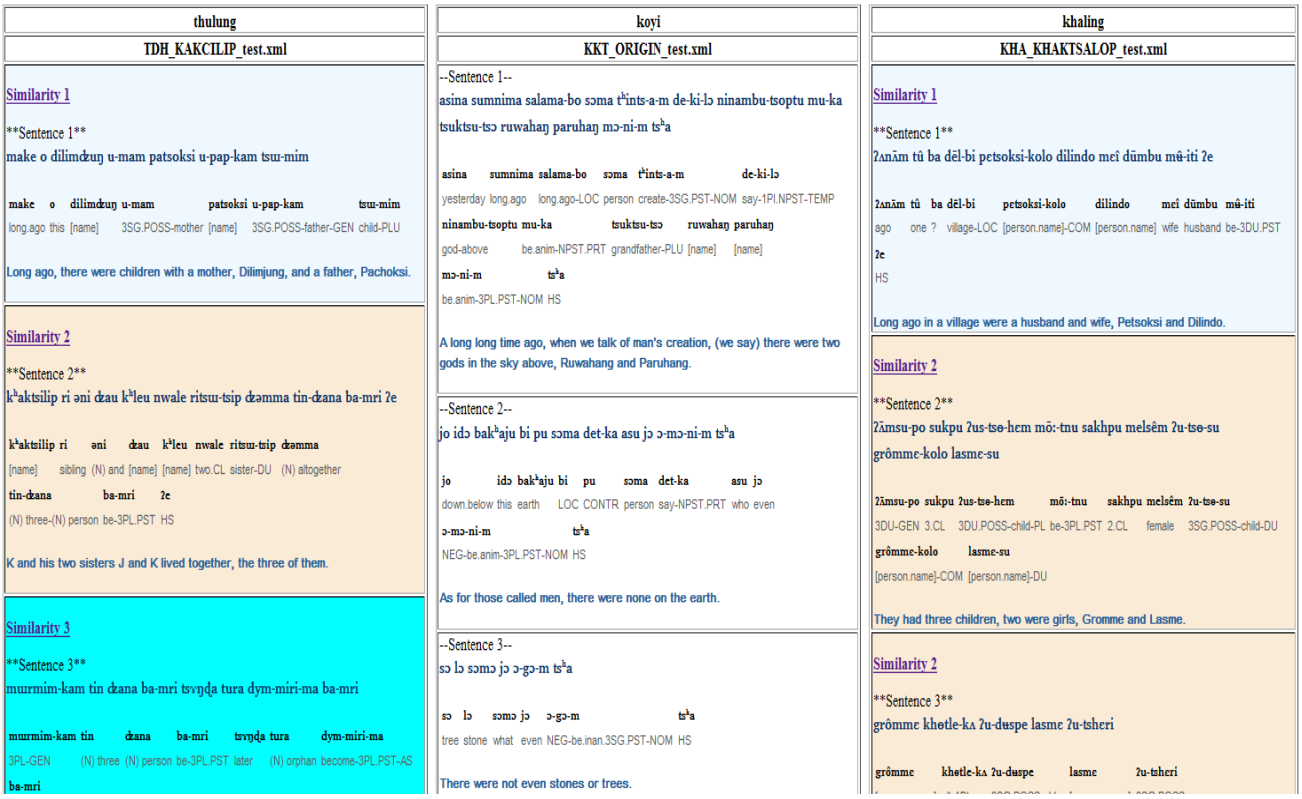
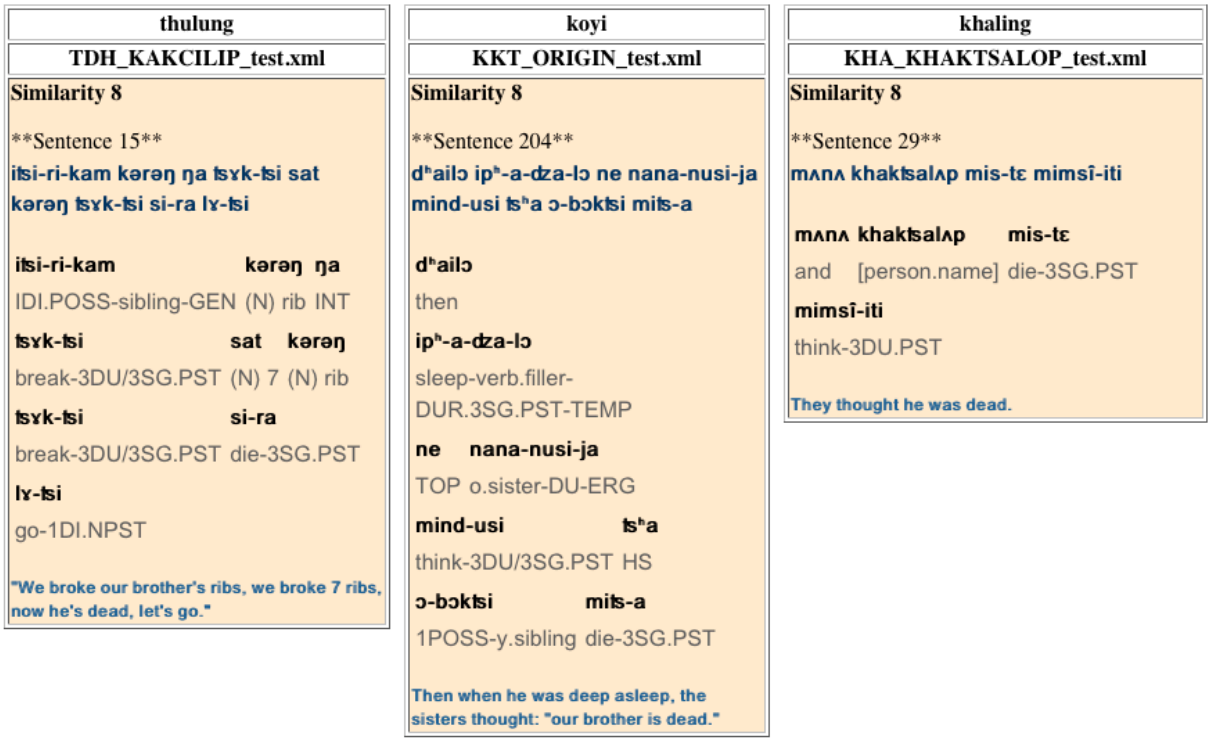


Figure 3. The "integral text view", with each language version of the story in its own column.



thulung	koyi	khaling
TDH_KAKCILIP_test.xml	KKT_ORIGIN_test.xml	KHA_KHAKTSALOP_test.xml
Similarity 8 **Sentence 15** **itsi-ri-kam kərəŋ ŋa tsɤk-tsi sat kərəŋ tsɤk-tsi si-ra lɤ-tsi** **itsi-ri-kam** **kərəŋ ŋa** IDI.POSS-sibling-GEN (N) rib INT **tsɤk-tsi** **sat kərəŋ** break-3DU/3SG.PST (N) 7 (N) rib **tsɤk-tsi** **si-ra** break-3DU/3SG.PST die-3SG.PST **lɤ-tsi** go-1DI.NPST "We broke our brother's ribs, we broke 7 ribs, now he's dead, let's go."	**Similarity 8** **Sentence 204** **dʰailɔ ipʰ-a-ʤa-lɔ ne nana-nusi-ja mind-usi tsʰa ɔ-bɔktsi mits-a** **dʰailɔ** then **ipʰ-a-ʤa-lɔ** sleep-verb.filler-DUR.3SG.PST-TEMP **ne** **nana-nusi-ja** TOP o.sister-DU-ERG **mind-usi** **tsʰa** think-3DU/3SG.PST HS **ɔ-bɔktsi** **mits-a** 1POSS-y.sibling die-3SG.PST Then when he was deep asleep, the sisters thought: "our brother is dead."	**Similarity 8** **Sentence 29** **mʌnʌ khaktsalʌp mis-tɛ mimsî-iti** **mʌnʌ khaktsalʌp** **mis-tɛ** and [person.name] die-3SG.PST **mimsî-iti** think-3DU.PST They thought he was dead.

Figure 4. View of one of the similarities across the three languages, the "similarity view".

Rechercher un terme :

OK

Texte	Phrase	Contexte gauche	Mot	Contexte droit	Gloses
TDH_KAKCILIP_test.xml	s2	ʦɯ mim kʰakʦilip ri əni ʤau kʰleu nwale	ritsɯ	ʦip ʤəmma tin ʤana ba mri ʔe mɯrmim	sister
TDH_KAKCILIP_test.xml	s60	m pəʦʰi meɖɖamma gana reʦa rak ta mɯ	ritsɯ	ʤeɖ ɖy kole ritsɯ bai ra m mɯ	sister
TDH_KAKCILIP_test.xml	s60	reʦa rak ta mɯ ritsɯ ʤeɖ ɖy kole	ritsɯ	bai ra m mɯ ʦɤrʦi lam al pa	sister
TDH_KAKCILIP_test.xml	s61	m mɯ ʦɤrʦi lam al pa luŋ ʦʰəhi	ritsɯ	dys ta memsaka by ry ʔe əni mɯram	sister
TDH_KAKCILIP_test.xml	s61	dys ta memsaka by ry ʔe əni mɯram	ritsɯ	ʦʰəhi ʤeɖ ɖy lo go make ŋa pəhila	sister
KKT_ORIGIN_test.xml	s127	munmuri kʰokʦ a kʰɔnʦ asi dʰanɔ sɔ pu	ʦʰekuma	nusi ʦʰoʔ si umnusi pu kʰur bi buwa	sister
KKT_ORIGIN_test.xml	s147	nɔ mɔ si lɔ ne aʤi ne ana	ʦʰekuma	nusi wɔ oko nɔ kʰur bi dʰoʔo kɔ	sister
KKT_ORIGIN_test.xml	s187	maʔa kim bi ne bɔkʦi ʦɔ mo ni	bigja	nɔ ʦi ʦɔ mo ni kʰɔlɔ mo ni	sister
KKT_ORIGIN_test.xml	s191	luʔ nɔ pʰiŋ u ʦʰa aʤi sɔma ʦɔ	bigjame	ʦɔ made mɔ ni m ʦʰɔ kim bi	sister
KKT_ORIGIN_test.xml	s278	lu ʦʰa naga ne hɔ inʦi sɔlɔ inʦi	ʦʰekuma	inʦi ʦi sɔlɔme ʦʰɔʔ ɔ kɔpɔ sɔma ʦɔ	sister
KKT_ORIGIN_test.xml	s334	nɔ ne heʔnɔ mɔ ʤa si ʦʰa ɔbɔ	ʦʰekuma	hoʔle dʰaiʔnɔ ŋɔ umbika sɔksu pu ana nuwa	sister
KKT_ORIGIN_test.xml	s345	ase papa luʔ si heʔŋɔ dʰam bi ne	ʦʰekuma	ʦi ʦʰɔm ka ʦʰuʔ mu ʦʰɔ ʦʰɔm ka	sister
KKT_ORIGIN_test.xml	s346	ʦʰɔm ka ʦʰuʔ mu ʦʰɔ ʦʰɔm ka ʦʰekumɔ	ʦʰekuma	ʦʰuʔ mu ʦʰɔʔ lɔ ne kɔpa mo a	sister
KKT_ORIGIN_test.xml	s347	ʦʰuʔ mu ʦʰɔʔ lɔ ne kɔpa mo a	ʦʰekuma	ʦi dja nɔ ʦjuri mu di pʰiŋ usi	sister
KKT_ORIGIN_test.xml	s350	se si ʦʰa ʦʰɔmdam bi ʦʰɔm ka ʦʰe	ʦʰekuma	ʦi ʦʰaŋgara pʰiŋ usi ʦʰa ʦʰaŋgara jɔ ja	sister
KKT_ORIGIN_test.xml	s353	a nɔ aʤi pi wa pu dja ʦʰa	ʦʰekuma	nɔ bo hɔbaʦ a nɔ kʰɔkʦulupa dja ʦʰa	sister
KKT_ORIGIN_test.xml	s362	si m ʦelbu nɔ kim hoʔle si dʰaiʔlɔ	bigjame	nusi ʦʰɔm ka bi pu risi pik uni	sister
KHA_KHAKTSALOP_test.xml	s125	ʔɛˆn ʤɵ jo ŋa mɛˆm ʦʉnʉ hɛm ʔɛ	bèŋmɛ	ʦɵ hɛm ʔɛ was ʦhoɔ¯m nu lo ʔathā:	sister

Figure 5. Concordance of the gloss "sister"