2006

Automated Adaptation Between Kiranti Languages

Daniel Richard McCloy
The University of Montana

AUTOMATED ADAPTATION BETWEEN KIRANTI LANGUAGES

By

Daniel Richard McCloy

B.S. Computer Science and Engineering, LeTourneau University, Longview, TX, 1991

Thesis

presented in partial fulfillment of the requirements
for the degree of

Master of Arts
in Linguistics

The University of Montana
Missoula, MT

Autumn 2006

Approved by:

Dr. David A. Strobel, Dean
Graduate School

Dr. Anthony Mattina, Chair
Linguistics Program

Dr. Mizuki Miyashita
Linguistics Program

Dr Nancy Mattina
The Writing Center

Dr. David E. Watters
Research Centre for Linguistic Typology

McCloy, Daniel, M.A., December 2006 Linguistics

Automated Adaptation Between Kiranti Languages

Chairperson: Dr. Anthony Mattina

Minority language communities that are seeking to develop their language may be hampered by a lack of vernacular materials. Large volumes of such materials may be available in a related language. Automated adaptation holds potential to enable these large volumes of materials to be efficiently translated into the resource-scarce language.

I describe a project to assess the feasibility of automatically adapting text between Limbu and Yamphu, two languages in Nepal's *Kiranti* grouping. The approaches taken—essentially a transfer-based system partially hybridized with a Kiranti-specific interlingua—are placed in the context of machine translation efforts world-wide.

A key principle embodied in this strategy is that adaptation can transcend the *structural* obstacles by taking advantage of *functional* commonalities. That is, what matters most for successful adaptation is that the languages "care about the same kinds of things." I examine various typological phenomena of these languages to assess this degree of functional commonality. I look at the types of features marked on the finite verb, case-marking systems, the encoding of vertical deixis, object-incorporated verbs, and nominalization issues.

As this Kiranti adaptation goal involves adaptation into multiple target languages, I also present a disambiguation strategy that ensures that the manual disambiguation performed for one target language is fed back into the system, such that the same disambiguation will not need to be performed again for other target languages.

ACKNOWLEDGEMENTS

It is a blessing to be able to work with people that you really enjoy, and in this regard I have been blessed at every turn.

This project could have never happened without the months of work poured into it by Marius Doornenbal and Jeffrey Webster, who have tirelessly dug to find answers and solutions, and coaxed or prodded me along when I needed it. –Thank you!

I cannot imagine how this thesis would have come to be but for Dr. Tony Mattina's guidance and his time invested in me over these years. I was already indebted to him for setting my wife on the path where I consequently met her. I am now doubly indebted. –Thank you!

It has a been a privilege, too, to receive much-needed help from such a master of Himalayan typology as Dr. David Watters. Not that I've grasped a fraction yet of what's really going on in these languages, but what little of it I've begun to understand is owing in great measure to him. –Thank you! Nowhere else will it be truer to emphasize that all errors in this paper are mine alone.

Dr. Mizuki Miyashita and Dr. Nancy Mattina have not only shouldered the task of serving on my committee, but have also invested time in my development and training. –Thank you!

Space does not permit me to name them all, but there have been many, many others, too, who have given of themselves to bring me to this stage. I am so grateful.

The greatest enabling and encouragement has come from my wife Adina, who has set aside her own desires time and time again to love me into reaching this milestone. –Thank you! How blessed I am indeed.

Soli Deo Gloria.

Dan McCloy
Missoula, Montana, November 2006

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

> "There are several major problems of minority languages in the modern society. In the age of globalization, there is a strong pressure to use a majority language everywhere, and although the democratic governments usually pay a great deal of attention to the needs of minorities, minority languages always are in danger of dissolving. One of the possible ways how to help to preserve a minority language might be using an MT [(Machine Translation)] system for producing relatively cheap translations from other languages, thus making available the texts which would not normally be translated."
>
> (Homola and Kuboň 2005)

Nepal is a land of about a hundred languages, the vast majority of which are endangered and scarce in resources. Nepali is the mother tongue of only about half the population, though, as the *lingua franca*, it has made enormous inroads into other-tongue communities. Until the revolution of 1990, minority languages were typically perceived by the government of Nepal to be a threat to national unity. Since that time, while there has been a growing acceptance of the value of the mother tongue, resources to aid in the development of these languages have continued to be scarce. Mother-tongue education and adult literacy programs have made some progress, but few materials are available in minority languages.

This is the story of a pilot study—undertaken by myself and two others—to assess how feasible it might be to automatically adapt text from one minority language to another. In particular, we examine the Kiranti languages of eastern Nepal, a cluster of languages known for their morphological complexity, which is overviewed in section 3. My specific focus is on the possibility for adaptation into the Yamphu language from the Limbu language.

We shall begin, however, by taking an overview of machine translation (section 2) to understand what it can do, and to identify the various strategies

that have been used: 'direct', 'transfer-based', 'interlingual', 'example-based', and 'statistical'. With an understanding of these strategies, and with an understanding of the implications of word-frequency issues (section 4), we can assess the suitability of the different strategies for the Kiranti situation.

An analysis of the Limbu and Yamphu languages begins in section 5 with structural comparisons of the finite verb. While abundant evidence of historical relationships between agreement morphemes can be found between the languages, complexity has crept into the diverging patterns. These morphemes can no longer be individually mapped directly between languages. However, it is evident that *in combination*, they convey the same functions across Kiranti languages. These languages are typologically unified in the *types* of functions marked on the verb: agreement for both agent and patient in the same eleven person-number combinations, tense/aspects encoded, and negation. We thus adopt an interlingua-like strategy of representing these functions abstractly in *functemes*, minimal units of function, in a "function-oriented" strategy described in section 6.

Section 7 tells the story of our attempts to implement this function-oriented strategy, first using the *Toolbox* program, and then switching to *CarlaStudio*. It also describes the "morphology" with which we implemented our hybrid interlingua.

Inherent in the transfer process are issues of ambiguity. Ambiguities arise wherever the source language does not make a distinction that the target languages do. These may be distinctions in form, such as those arising from coincidental homophony in the source language. Such ambiguities are inherent in the source language analysis. For example, an analysis of the English word *'bank'* involves an ambiguity that includes both nouns—some of which might be glossed *'river shore*' and *'money store*'—and verbs, including one that might be glossed *'to tilt while turning.'* Alternatively, ambiguities may arise where the target language grammaticalizes a semantic distinction

not present in the grammar of the source language, such as a distinction between dual and plural number. Some of these ambiguities can be resolved automatically during the adaptation process, while others require manual disambiguation. Section 8 discusses the disambiguation process, and an innovative 'disambiguation feedback process' that should be useful in a multi-target translation system.

Having established an architecture for a Kiranti function-oriented approach and for a disambiguation cycle, we turn to consider other issues for adaptation. Section 9 is primarily a comparison of some of the major typological features of Limbu and Yamphu. The nominal typology includes a comparison of the case-marking systems, and also of the systems for marking vertical deictics, that is, the vertical height of a referent relative to the deictic center of the speech act. The verbal typology addresses the phenomenon of object incorporation which results in prefixes apparently being inserted in the verb stem. Another highly significant issue is that of nominalization, a phenomenon that is central to Kiranti languages, applying at all levels, but with seemingly varied semantic effect. Also briefly addressed are a selection of issues of information structuring, including sequencing, expressing of causal relationship, and various other markings on the clause. Then, to obtain a sampling of issues that may best come to light outside the framework of typological issues, we examine parallel Limbu and Yamphu texts, and draw further inferences for adaptation.

Finally, in section 10, we are able to step back and assess the findings. Essentially, it is clear that the function-oriented strategy that was adopted provides a means to extend the reach of automated adaptation between Limbu and Yamphu, but it remains to be seen whether "further" is "far enough". Steps are outlined, however, requiring the participation of a Yamphu speaker, as to what it would take to proceed further with this pilot

study, in the hope that adaptation between Kiranti languages may ultimately prove fruitful for the Kiranti language communities.

2. OVERVIEW OF MACHINE TRANSLATION

2.1. The Capabilities of Machine Translation

This study is part of the field traditionally referred to as *machine translation* (MT), the process of programmatically transforming text from one language variety to another.

A typical starting point for investigating the abilities of machine translation is to try out the online translation tools. Amusement seekers may get text translated into and back out of multiple languages, but in all fairness, the tools should be evaluated on the basis of a single transfer pass. These online systems are often called *gisting* tools, as their output is usually good for giving the gist of the source text.

There are several web-sites offering online translation, but almost all of them provide the same translation generated by Systran. The most well-known are *BabelFish* at the AltaVista site, and *Google Translate*. (Google also adds some enhancements to the Systran results, and uses its own engine for some additional languages, such as Arabic.) The only completely non-Systran translation server I could find on the web was IBM's *WebSphere Translation* server.

For a sample of the capabilities of these two servers, compare their translations of a brief Spanish-language news article (in which "retain an actuary" does not mean "hire an insurance statistician", as might be imagined).

(1) Output of Two Translation Servers

Systran Translation	**IBM WebSphere Translation**
Inhabitants of Santa Catarina in Tláhuac retain to actuary and patrol	**Inhabitants of Saint Catarina in Tláhuac retain actuary and patrol**
Wednesday 6 of September of 2006 *Hour of publication: 09:46*	Miercoles *6 September 2006* *Time of publication: 09:46*
Inhabitants of the town of Santa Catarina, Tláhuac delegation, retained from early hour a patrol and to an actuary who was going to carry out an evacuation, reason why demand the presence of authorities.	Inhabitants of the Holy Catarina, Tláhuac delegation, village authority patrol and an actuary who was going to effect an ejection, by what they require her witnesses retained from temprana hour.
The Secretariat of Seguridad Pública (SSP) of the Federal District informed that to the place already Gabriel Regino goes to the undersecretary to speak with the inconformes and to treat that the situation is standardized.	Federal District Public Safety Secretariat (SSP) informed that to the place one already the undersecretary manages Gabriel Regino to talk with the nonconformist ones and try that the situation normalizes one.
This morning in the well-known zone like the Cross, of the town of Santa Catarina, appeared an actuary to make an evacuation and when not allowing it to it around one hundred people the presence of the public force was requested.	An actuary introduced this morning in the known zone as Saint's village Catarina Cross, to make an ejection and when not allowing it to them the police presence asked for around a hundred people.
Therefore they arrived elements of the body of uniformed grenadiers and; nevertheless, the inhabitants seized of patrol AC019 and stopped the actuary who was going to fulfill his work.	By this reason grenadiers' body elements arrived and standarized; however, the inhabitants authorized of the patrol AC019 and one arrested the actuary who was going to fulfil his work.
Before that situation one asked for the presence of more grenadiers and the one of the undersecretary of Public Security inhabitant of the capital, Gabriel Regino, who already goes to the zone to engage in a dialog with the inconformes.	Before that situation Gabriel Regino who already goes to the zone to talk to the nonconformist ones capitalina, asked for the presence of more grenadiers and that of the Public Safety undersecretary.

Even between the two renderings, deciphering what happened is not necessarily clear. (The source text and a human translation are available in Appendix A.) Neither translation server is clearly superior to the other, although the IBM WebSphere Translation server apparently incorporates some statistical techniques (though evidently not a spelling check). The main

point is that how successful these translations are perceived to be still depends to a great extent upon the level of the user's expectations: the general gist versus publishable quality.

It should also be noted that these samples do not represent the highest quality of which machine translation is capable. In the 2005 Machine Translation evaluation by the National Institute of Standards and Technology (NIST), an agency of the U.S. Commerce Department's Technology Administration, Google Labs—among several participating organizations—generated the best Arabic to English translation, but each sentence produced took on average forty hours of processing time. (With their large server bank, they are able to tackle greater statistical processing. Obviously, the depth of processing performed in this contest was not comparable with what they offer via their free web translation server.)

Moreover, if the translation is geared for a limited semantic/usage domain, quality can be improved. For example, Canada's Météo system has translated English weather bulletins into French every day since 1977, and its success is due in no small part to the limited domain of its use. For Météo, the word *"front"* is always a noun meaning *"weather front"*, never any of the other meanings that *front* can have in English.

Language Weaver is a company with a reputation for being on the cutting-edge of Arabic-to-English translation, a field in which U.S. government agencies have developed a high interest in recent years. Language Weaver does not provide free translation demonstrations, but they do advertise some selected samples of what they can produce, and these are clearly of higher quality than the translation of Spanish in (1).

(2) Sample Arabic translation by Language Weaver

Original text:	بغداد 15-1 (اف ب) - اعلن متحدث باسم وزارة الخارجية العراقية ان قيام المفتشين الدوليين بزيارة حي قريب من قصر رئاسي في بغداد يشكل "خطوة استفزازية".
Machine translation:	Baghdad 1-15 (AFP)-- announced a spokesman for the Iraqi Foreign Ministry that the international inspectors to visit district near the presidential palace in Baghdad as a "provocative step."
Human translation:	Baghdad 1-15 (AFP) - A spokesman for the Iraqi Foreign Ministry stated that the visit by the international inspectors to a district close to a presidential palace in Baghdad was "a provocative step".

These cutting-edge results were achieved by a complex statistics-based system with advanced linguistic modeling, utilizing large corpora of parallel texts, and developed through years of labor by brilliant people with substantial funding.

2.2. Types of MT Strategies

The field of machine translation has come a long way. Warren Weaver, a statistician who had been overseeing American cryptography in World War II, is generally credited with the idea that digital computers might be programmed to automatically translate between natural languages. As a result of his proposals, a variety of MT projects sprang up in the US and around the world, and the field of machine translation was off to an enthusiastic start by the early 1950s. We will examine the four basic types of strategy employed since that time.

2.2.1. Direct Translation

Direct Translation is a strategy of mapping the source language directly to the target language. One of the first such projects was Georgetown University's GAT system for translating Russian to English, started in 1952. Two years later, with a grammar of just six rules and a vocabulary of 250 words, it demonstrated the ability to translate 49 hand-picked sentences from Russian. After another ten years of development funded by the U.S.

government, it was installed at the Oak Ridge National Laboratory, where it was used for many years to translate Russian physics journals (Slocum 1985). The quality of GAT's translation was poor, but in the aftermath of the *Sputnik* program, the U.S. was desperately scrambling to catch up with Soviet advances in physics, so a poor translation was better than not having any translation. Indeed, the *Sputnik* launch itself

> "was perceived as a drubbing not only of American rocket science, but of American intelligence gathering, *hampered by a lack of rapid means of translation*. (Months before the liftoff, a Soviet hobbyist magazine alerted ham-radio enthusiasts to the imminent launch of an experimental satellite, even providing a shortwave frequency for tracking it. The US Navy, however, never saw a translation of the article. After the launch, it scrambled for days to reconfigure its 'radio fence' to intercept *Sputnik's* transmissions and figure out what it was doing.) Edward Teller, the father of the hydrogen bomb, declared shortly after the launch that the US had lost 'a battle more important and greater than Pearl Harbor.'"
>
> (Silberman 2000, emphasis added)

Other governments similarly funded MT research. By 1962, there were MT projects in the U.S., the U.K., Italy, Japan, the U.S.S.R., China, Mexico, Belgium, Yugoslavia, Hungary, East Germany, and France (Silberman 2000).

One of Weaver's proposals in 1949, however, had been that if language could be reduced to universals, these would form a means by which computers could perform translation better than by the "direct" route:

> "Think, by analogy, of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another, they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed. But, when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers.
>
> Thus it may be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is *not to attempt the direct route*, shouting from tower to tower. Perhaps the way is to descend, from

each language, down to the common base of human communication- the real but as yet *undiscovered universal language* and then re-emerge by whatever particular route is convenient."

(Weaver 1949, emphasis added)

2.2.2. Translation via Interlingua

The second type of machine translation strategy implemented Weaver's "universal language" philosophy as an *interlingua*, an idealized unambiguous semantic representation derived from the source text, and from which text in the target language could be generated. In some systems, this representation has been based on a real human language, such as Esperanto or, in case of the ATAMIRA system in the 1980s, the South American language Aymara. In other systems, the interlingua has been much more abstract. For example, the Universal Networking Language (UNL) project (1996 to the present) of the United Nations University in Tokyo employs a representation that aims to be completely language-neutral, except that instead of going so far as to assign numbers to represent their "universal words", assigned English labels are used:

Figure 1 Simplified UNL representation of "*The small car is not red.*"

```
attrib( red.@present.@not.@topnode, car.@def.@topic )
attrib( small, car )
```

(Simplified from Hong and Streiter 1999)

In this example, the first line specifies a relationship between *red* and *car*, while the second line specifies a relationship between *small* and *car*. The '@' operator specifies additional semantic/functional information. That both lines refer to the identical universal word *car* links them as being co-referential within a combined predication. (If they referred to different cars, additional marking would have been employed.) Of particular importance is the '@topnode' label, which specifies that what is being predicated is the '*not red'*-ness, not the *small*ness. Thus, the English generated from this UNL cannot be *'The car which is not red is small.'*

Possibly the most productive interlingua-based system in use commercially today is KANT, a system used by Caterpillar for translating the manuals for their earth-moving equipment into several (primarily European) languages. Here is an example of KANT's interlingual representation of the sentence: *"The default rate remained close to zero during this time."*

Figure 2 Sample KANT Interlingua

```
(*A›REMAIN                                  ; action rep for 'remain'
.  (FORM FINITE)
.  (TENSE PAST)
.  (MOOD DECLARATIVE)
.  (PUNCTUATION PERIOD)
.  (IMPERSONAL ›)                           ; passive + expletive subject
.  (ARGUMENT›CLASS THEME+PREDICATE)         ; predicate argument structure
.  (Q›MODIFIER                              ; PP semrole (generic)
.  .  (*K›DURING                            ; PP interlingua
.  .  .  (POSITION FINAL)                   ; clue for translation
.  .  .  (OBJECT                            ; PP object semrole
.  .  .  .  (*O›TIME                        ; object rep for 'time'
.  .  .  .  .  (UNIT ›)
.  .  .  .  .  (NUMBER SINGULAR)
.  .  .  .  .  (REFERENCE DEFINITE)
.  .  .  .  .  (DISTANCE NEAR)
.  .  .  .  .  (PERSON THIRD)))))
.  (THEME                                   ; object semrole
.  .  (*O›DEFAULT›RATE                      ; object rep for 'default rate'
.  .  .  (PERSON THIRD)
.  .  .  (UNIT ›)
.  .  .  (NUMBER SINGULAR)
.  .  .  (REFERENCE DEFINITE)))
.  (PREDICATE                               ; adjective phrase semrole
.  .  (*P›CLOSE                             ; property rep for 'closer'
.  .  (DEGREE POSITIVE)
.  .  (Q›MODIFIER
.  .  .  (*K›TO
.  .  .  .  (OBJECT
.  .  .  .  (*O›ZERO
.  .  .  .  .  (UNIT ›)
.  .  .  .  .  (NUMBER SINGULAR)
.  .  .  .  .  (REFERENCE NO›REFERENCE)
.  .  .  .  .  (PERSON THIRD))))))))
```

(Czuba, Mitamura and Nyberg 1998)

The KANT system was developed in conjunction with Carnegie Mellon University, and it is a *knowledge-based* interlingual system, in that

additional properties of words are utilized for achieving disambiguation. Essentially, this addresses the issue that caused Bar-Hillel to abandon machine translation in 1959 (Hutchins 1999). His oft-cited example is that machine translation could never properly translate the sentence: *"The box was in the pen",* as in the context: "Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy." Bar-Hillel argued that no existing or imaginable program would enable an electronic computer to determine the appropriate sense of the word *pen*—whether *playpen* or *writing instrument*—in the given sentence within the given context. For a computer to be able to tell the difference, it would need not just a dictionary but also "a universal encyclopedia" (Bar-Hillel 1959).

Knowledge-based systems seek to provide the necessary encyclopedia of real-world knowledge that can constrain interpretation. For example, one sense of the adjective *light* is 'not heavy', while another is 'not dark'. If the computer can be made to recognize that in a given context, concepts relating to *shade/color* are more significant than concepts relating to *weight*, on this basis the correct "universal word" can be selected. In Bar-Hillel's contrived example, the sense of *playpen* is likely only in contexts relating to children, and the word *toy* does indeed introduce that context, so on this basis, a system may be able to make the correct judgment. In more advanced knowledge-based systems, spatial concepts may also result in other *selectional restrictions* that constrain larger items from being inside a smaller item, such as something in the size range of a "toy box" being inside something in the size range of a writing instrument. Selectional restrictions are also used in resolving ambiguities in speech-to-text processing, such as saying that a verb like *swallow* requires an animate being in the agent role and physical object in the patient role. However, consider these metaphorical uses: *"I swallowed his story, hook, line, and sinker",* and *"The supernova swallowed the planet"* (Manning and Schütze 1999). Thus, the use of

metaphor continues to be a problem for knowledge-based systems. For interlingua systems in general, the greatest problem has proven to be that a failure to get an analysis results in zero output, as the interlingua cannot represent what the system cannot analyze.

2.2.3. Transfer System

The third basic type of strategy was the *transfer* system. Whereas the interlingua approach's search for universal semantic unity was a lofty ambition, the transfer system traded some theoretical elegance for a practical and robust strategy. It separated out the processes into three distinct stages:

- **Analysis** of Source Language (SL) in terms of its own grammatical structures.
- **Transfer** of SL structures to the structures of a particular Target Language (TL).
- **Synthesis** of TL structures into TL surface forms.

This provided the modularity that the direct approach had been missing. The direct approach had typically relied on a single bilingual dictionary, and on rules to transform source elements into final target elements. Virtually none of this could be reused in a parallel project to translate from a different source language or to a different target language. In contrast, with the transfer approach, the analysis of the source language is performed without any consideration of target language structures. The analysis is often based on a particular linguistic theoretical framework, such as dependency grammar or categorial grammar. The second stage, the transfer stage, is transfer specifically between a source language and target language pair. The final stage, synthesis, applies for a target language regardless of which source language the text originated in. In contrast with the interlingua approach, text that could not be fully analyzed at least does not result in zero output. Transfer systems to this day continue to form the basis of most

commercial MT systems. Possibly the best-known transfer system on the market is that of Systran, originating in the early 1970s.

2.2.4. Corpus-based Methods

The fourth basic type of translation strategy has only really taken off since the late 1980s and early 1990s. These strategies are actually of various types that fall into the category of *corpus-based* systems, taking an empiricist approach, and requiring large corpora of parallel texts. The two most significant corpus-based systems are called *example-based translation* and *statistical machine translation.*

Example-based Translation

Example-based translation was first suggested by Nagao Makoto of Kyoto University as a means to achieve high-quality translation. Translation is essentially done by analogy, re-using portions of similarly translated text. For example, suppose an aligned Japanese-English bilingual corpus contains these two pairs:[1]

(3) **He buys a notebook.**
Kare ha nouto wo kau.

(4) *I read a book on international politics.*
Watashi ha kokusaiseiji nitsuite kakareta hon wo yomu.

Based on these, the English sentence:

(5) **He buys** *a book on international politics.*

can be translated into Japanese as:

(6) **Kare ha** *kokusaiseiji nitsuite kakareta hon* **wo kau**.

(Sato and Nagao 1990)

Of course, to find exact matches, this requires huge corpora, so it is often hybridized with other systems to achieve "fuzzy matching" of similar but not identical clauses.

[1] The paper from which this example is quoted apparently uses a Japanese transcription scheme that is not entirely phonetic.

This is essentially the philosophy that underlies *Translation Memory* (TM) systems, though generally in a less automatic way. For the purposes of high-quality translation of unconstrained source material, many professional translators distain to 'post-edit' (that is, manually correct or polish) the results of machine translation, as the types of errors that are made can be awkward, compared to post-editing the translation of a junior human translator. TM systems allow a translator to "recycle" the translation he previously made for a word, phrase, or clause, or to do likewise based on the prior translations of his colleagues. The leading product in this market would currently seem to be TRADOS from SDL International. The "corpus" in TM systems like this is only built as the translator translates, but as the domain is typically quite constrained, it evidently works well enough to sell.

Statistical Machine Translation (SMT)

At the cutting edge of machine translation today are statistical techniques. There has been an explosion in research in statistical techniques, starting with IBM's work in the late 1980s, as formulated in Brown et al. 1993. Interestingly, this had been a proposal of Warren Weaver himself right back at the start of machine translation:

> One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.".
>
> (Weaver 1947)

Indeed, there had been some statistical attempts in the 1950s, generally referred to as "brute force" (as opposed to the "perfectionist" interlingua and transfer systems), but the particular statistical technique that now began to bear fruit was the "noisy channel" model being utilized in speech-to-text processing.

This model takes Weaver's hypothesis quite literally, as if Russian is actually garbled English. For having such an obviously untrue basis, the results nonetheless speak for themselves. The basic concept behind it is that while it may be extremely difficult to directly assign a probability that a given English (i.e. target-language) sentence is the most likely translation of a given Russian (i.e. source-language) sentence, by using Bayes Rule, we can calculate that probability reasonably well by first calculating the probability that the Russian sentence is the translation of the English, and then multiplying that by the probability that the English is *good* English. The component that calculates the probability of the Russian being a translation of the English is called the *translation model*, and it is derived from parallel Russian-English corpora. The component that calculates the probability that a given English sentence is good English is called the *language model*, and it is derived from an even larger English corpus. The translation model doesn't have to be all that good, because the language model factors out the improbable English sentences. It only needs to say how well an "English bag of words" corresponds to a "Russian bag of words". Separately, the language model will take care of whether the English word order is good.

How, then, does the language model calculate the likelihood that the English is well-formed? The traditional SMT approach is to ignore grammar, and come up with a figure based on the probability of the words of the sentence appearing in their given sequence. Of course, there are many valid sequences of words that will be completely unattested in the corpus. Chomsky argued that statistical models would fail, because unattested utterances would be assigned the same zero probability as ungrammatical ones" (Manning and Schütze 1999). What SMT does, however, is to look at smaller sub-sequences, typically tri-grams (sequences of three words) or bi-grams (sequences of two words). If a string contains a lot of reasonable tri-grams and bi-grams (more generally called *N-grams*), it has a higher probability of being well-formed.

Consider the heretofore unattested[2] sentence, *"Sorry I kicked your cat."* A bi-gram language model determines the probability of this being a good English sentence by looking at each sequence of two words. What's the likelihood of the word '*sorry'* being the first word of a sentence? What's the likelihood of it being immediately followed by the word *'I'?* What's the chance of *'kicked'* following *'I'?* And so forth till the end of the sentence.

Formally, if we let **b**(*y* | *x*) represent the probability that word *y* follows *x* in the corpus, the probability that the entire sentence is valid can be computed as:

P(*sorry I kicked your cat*) ~
  b(*sorry* | <start-of-sentence>)
x **b**(*I* | *sorry*)
x **b**(*kicked* | *I*) *
x **b**(*your* | *kicked*) *
x **b**(*cat* | *your*) *
x **b**(<end-of-sentence> | *cat*)

Applying the same process to other combinations of words will result in lower probabilities for proposed word orderings such as *"cat your kicked I sorry"* or even *"sorry your cat kicked I".* Actually, the word order *"I kicked your sorry cat"* is grammatical, so it should probably result in a relatively high ranking. Will it outrank our original sentence? If your monolingual Russian neighbor uses this sentence when you expect he's trying to apologize, perhaps he's using SMT software based on an English corpus in which apologies are statistically rare. It seems to be a weakness of SMT that a frequent use will outrank an infrequent use. (Of course, its proponents say that that is precisely the strength of it too, considering the errors that rule-based systems are prone to, treating all ambiguities as if they were equally likely.)

[2] Well, I don't think I've ever heard anyone say that, and nor does it show up in a Google phrase search of either the Web or Google's book library. Neither does the similar apology, "Sorry that I kicked your cat."

In theory, the answer to this problem is that even though a sentence may be assigned a high probability by the language model, if the translation model assigns it a low probability of corresponding to the source, it is less likely to ultimately outrank other sentences. However, consider this: If results of a Google Web search[3] are reasonable indicators, the word *"sorry"* appears about 440 million times on the Web. The sequence *"sorry I"* appears 417 million times, and the sequence *"sorry cat"* appears just 12,700 times. That is, the word *"sorry"* is followed by the word *"I"* 95% of the time, or is followed by the word "cat" less than 0.003% of the time. In other words, the sequence *"sorry I"* is about 33 thousand times more likely that the sequence "sorry cat". Thus, it is difficult for me to imagine how a hypothesis of *"I kicked your sorry cat"* could ever outrank the hypothesis of *"Sorry I kicked your cat,"* regardless of which one is the proper translation of the source text. Supposing a situation in which the *"I kicked your sorry cat"* hypothesis was indeed the proper translation, this hypothesis should thus be assigned a somewhat higher *translation*-model factor than the other, but not a dramatically higher factor, as both hypotheses represent the right "bag of words" (in this case relating to kicking and cats). The *language*-model factor that each receives, however, would seem to overwhelm the significance of the translation-model factor, as the wrong hypothesis contains a sequence occurring 33 thousand times more frequently than the sequence in the right hypothesis.

Another aspect of this model that seems problematic to me is the treatment of sentences in which the agreement is not close together. For example, consider the verb *eats* in the sentence: *"My friend with three cows, two ewes, and half a dozen goats <u>eats</u> a lot of cheese."* An N-gram language model is going to prefer to use the word *"eat"* there instead, as the bi-gram *"goats eat"* is statistically more probable than the bi-gram *"goats eats"*. Perhaps SMT sometimes gets away with this working right because the translation model

[3] Google search results here are filtered by Google's SafeSearch feature, so they do not include statistics from the dark side of the Web.

insists on ranking *eats* as the more probable word to correspond to the source word, which also happens to show third person singular agreement. But if the source language didn't have verbal agreement, it seems the translation model would have nothing to outweigh the language-model's tendency to prefer the more probable bi-gram *"goats eat"*.

As you can see, neither the translation model nor the language model have the tiniest whit of linguistic knowledge. Indeed, in the early days of SMT, the attitude seemed to be, "Who needs linguists any more?" It seemed that even semantic issues could resolve themselves empirically, because, in the words J. R. Firth back in 1957, "You shall know a word by the company it keeps."

And yet, in the last few years, the cutting edge of SMT has begun to incorporate greater levels of linguistic modeling (cf. Koehn and Knight 2003, Charniak, Knight and Yamada 2003), with the recognition that if lousy models can give understandable output, good models should be able to give that much better of output. Indeed, this mixing of strategies is occurring throughout the field, with transfer systems, interlingua systems, and statistical systems borrowing ideas from each other in order to most effectively deal with the task before them.

2.3. Recent MT innovations in other resource-scarce situations

Since the major advances in machine translation in the last decade have been corpus-based, it's worth considering whether there are ways in which these advances can be applied to the smaller languages, where resources such as parallel corpora are scarce. There are some interesting studies that have looked at techniques of tackling resource-scarce MT.

Al-Onaizan et al. (2003) describe such a study at the Information Sciences Institute (ISI) at the University of Southern California. ISI is at the forefront of SMT innovations that are typically dependent on huge resources, such as

million-word parallel corpora, but this study acknowledges that "for most language pairs, [corpus] data is scarce, and current [SMT] techniques do not work well." Human translators, on the other hand, knowing nothing of the source language except what can be inferred from a limited bilingual text, are somehow able to translate dramatically better than any SMT system trained on the same text. This study aimed to identify what strategies the human translators were using, and to see if any of these could be incorporated into an SMT system to improve the quality of the results.

The experiment involved the translation to English from Tetun, a language of East Timor. Participants were provided with a bilingual corpus containing about a thousand aligned sentence pairs. For a separate ten-sentence news article, only the Tetun text was provided. The challenge for the participants was to generate the English translation of these Tetun sentences.

Some of the participants did, indeed, manage to produce translations that were strikingly similar to the reference translation. A variety of techniques were used, including: a process of elimination by which certain words could be identified; a recognition of cognates (e.g. *grupu–group* and *diskasaun–discussion*); a pre-processing deletion of certain high-frequency grammatical function words, and a determination meaning from a variety of translations (e.g. from the various translations of *presiza—needed*, *necessary*, *need to*, *required*, *have to*—the concept of *necessity* can be inferred). Some of the strategies are based upon real-world knowledge of what possible translations would make coherent sense. (Naturally, the quality of the results were also significantly boosted by the participants' inbuilt English language model being far superior to the best computational model ever built.) However, some strategies were such that it might be possible to implement them in a computational approach. For example, software that can derive the concept of *necessity* from the various words by which it is translated could then go on to select the optimal expression for *necessity* according to the given context.

This would actually be an interlingua-like strategy applied within a statistical framework. Another computationally feasible strategy is that of probabilistic cognate recognition, the resulting hypotheses of which can feed the process of elimination. One strategy that was proven effective in improving the quality of the SMT-generated text was that of deleting at the outset any high-frequency function words that have no corresponding word in the target language, English. This same strategy has been similarly implemented in places in our Kiranti project, as discussed in section 9.

Turning from conceptual studies to actual applications of resource-scare translation, Lavie et al. (2004) present a Carnegie Mellon University project for a "trainable transfer-based machine translation approach for languages with limited resources". It demonstrates an approach for Hindi to English MT, in which a specially-elicited corpus of two thousand phrases and sentences form the basis from which the system can automatically generate transfer rules. I wonder if calling this a "transfer-based" system is perhaps a misnomer, as it seems to be more of a "direct translation" model that has been fitted into an SMT framework. It is direct in that the rules are not separated into analysis, transfer, and synthesis. It is an SMT framework in that the generated rules form the basis for a translation model, generating a lattice of options, while the language model, being English, is the SMT standard, based on the enormous corpus available in English. For comparison purposes, paralleling the automatically-learned transfer rules, an alternate set of rules were manually written according to a knowledge of Hindi linguistics. The system was tested with sentences being generated by the following processes:

- **Standard SMT Only**: This resulted in the lowest quality results.
- **"No Grammar"**: This took advantage of word-to-word and phrase-to-phrase transfer rules, similar to a bilingual dictionary, but not of the

syntactic transfer rules. This resulted in significantly better performance than the SMT Only approach.

- **"Learned Grammar":** This utilized the automatically-learned syntactic transfer rules. It performed only marginally better than the "no grammar" approach.

- **"Manual Grammar":** Using the manually-written rules, it significantly out-performed the "learned grammar" approach.

- **SMT + Manual Grammar:** Combining the possibilities generated by the SMT and Manual systems, and submitting these to the language model resulted in score a bit better than that of the manual grammar alone.

From this I conclude that while the "learned grammar" approach is perhaps more suited to larger corpora, statistical techniques can improve the quality of manually-written rules to transfer from resource-scarce languages to resource-rich ones. However, this approach will unfortunately not help much for transfer where the resource-scarce language is the target, as it is the target language model—not the translation model—that requires the substantial resources.

Research or development for machine translation of any kind into a minority language is rare. There are strong commercial incentives that motivate translation between Japanese and English. For European languages, the needs of governance add additional impetus, while for Arabic-to-English translation, it is presumably the intelligence priorities of the U.S. government (reflected in their funding) that has driven the recent expansion of MT focus to Arabic. (Chinese-to-English is the other competition in which organizations participate in the annual NIST evaluations.) Those same incentives do not exist for languages spoken by populations numbering in the mere thousands. Research and development of machine translation strategies for such languages is rare indeed.

Homola and Kuboň (2005), however, present an approach for machine translation into a minority language: Lower Sorbian, spoken in Germany. Its genetically closest "major language" relative is Czech, and thus their study was based on using Czech as the source language. Due to the high degree of structural similarity, the system's architecture could be transfer-based and relatively simple. This study underscores the value of a transfer-based strategy for adaptation between closely-related languages.

2.4. History of MT in South Asia

By far the most ambitious efforts in machine translation in India are those of the AnglaBharti system being developed by the Indian Institute of Technology in Kanpur, under the leadership of R.M.K. Sinha. The program is supported by the Technology Development for Indian Language (TDIL) program of the Government of India (Sinha 2003).

AnglaBharti is a general foundation for Machine-Assisted Translation (MAT) from English to various Indian languages. MAT is distinct from MT in that greater emphasis or recognition is given to the role of the human editor. Built on top of this foundation are language-specific systems, such as AnglaHindi which takes the AnglaBharti output, and from it generates a Hindi draft, which is then manually post-edited into "good" Hindi.

This system incorporates elements of a wide variety of strategies: interlingua, syntactic transfer, example-based translation, knowledge-based selectional restraints, and even some statistical elements. Essentially, AnglaBharti adapts English text into an interlingual form named *Pseudo Lingua for Indian Languages* (PLIL). As the languages of India belong to four different language families (Indo-Aryan, Dravidian, Austro-Asiatic, and Tibeto-Burman), AnglaBharti is designed to generate four corresponding "flavors" of PLIL. Each of these four interlingual forms is designed according to the typological needs of the language family. The Indo-Aryan interlingua is thus

Sanskrit-oriented, such that Paninian grammar (together with statistical and example-based techniques) can generate Hindi or other Indo-Aryan text.

AnglaBharti is a relevant model for our multi-language strategy (discussed in section 3.3). This model, however, requires manual disambiguation to be repeated in each target language. Perhaps an adaptation of our disambiguation feedback cycle (described in section 8.3) could enhance the AnglaBharti architecture.

According to Roa (2001), a handful of other projects have sprung up in India, with roots that reach back to the late 1980s or early 1990s. These include the Anusaaraka project, which is not focused on machine translation *per se*, but rather on using principles of Paninian grammar to map words into Hindi from various languages of South Asia, including not only close Indo-Aryan relatives such as Marwari and Punjabi, but also Dravidian languages such as Telegu and Kannada. This system has mainly been applied for children's stories. Like AnglaBharti, this project began at IIT Kanpur, but it later moved to the Centre for Applied Linguistics and Translation Studies (CALTS) at the University of Hyderabad.

As for machine translation efforts specifically in Nepal, the first such project (as far as I am aware) was by Watters during the period 1986 to 1992. His work focused on inter-dialectal adaptation in the Kham language of Western Nepal, transferring from the Takale dialect to the Ghamale dialect. These language varieties have much in common, as indicated by lexical similarity counts of up to 96%. However, such "cognates" are not necessarily recognizable to speakers of the dialects, as a number of systematic changes have occurred in each dialect since the time that Proto-Kham began to diverge. As a result, mutual intelligibility between these two "dialects" is down in the mid 30% range (Watters 1988).

A strategy for morpheme parsing (i.e. for segmentation of a word into its constituent morphemes) that Watters nicknamed the "jitterbug scanner" has

proven valuable for Kiranti analysis, too. In contrast to languages in which derivational and inflectional affixation work by first attaching each of the prefixes, and afterwards attaching each of the suffixes, or vice versa, Watters found that in Kham, the order of affixation could alternate between prefixes and suffixes. The significance of this is that one of the key techniques for morpheme-parsing depends on *category mapping*, the constraint that any given affix can only attach to one or more specific categories (such as noun, intransitive verb stem, inflected verb, etc.), and that its affixation optionally results in a change of category.

Some background may be helpful here: The parsing process is basically a matter of trying to produce a list of all of the possible combinations of morphemes (allomorphs, actually) that could be strung together to make the word we are trying to parse. For example, the English word *extradition* could be parsed as either *ex* + *tradition* or *extradit* + *ion*. Likewise, the English word *detonatable* can be hypothesized to contain English morphemes *de.ton.a.table* or *de.ton.at.able* or *detonat.able*, among other possibilities. Category mapping is one of three strategies for eliminating bad parses during the analysis. The other two are the use of *orderclass constraints,* based on a "slot and filler" view of morphology such that each morpheme is assigned a numeric order class, and *morpheme co-occurrence constraints*, by which certain combinations of morphemes may be accepted or rejected. e.g. In Caquinte, the future prefix can only be present if the future suffix is also realized in the word (Black and Black 2005). If all of these analysis strategies have done all they can to reduce bad parses and yet multiple options remain, the ambiguity gets passed on for possible disambiguation based on syntax, and if unresolved at that stage, manual disambiguation will ultimately be required. This may be the case when multiple parsings are valid. e.g. German *wachtraum* could be either parsed as *wach* + *traum 'day dream'* or as *wacht* + *raum 'guard room'* (Hutchins 2003).

During analysis, then, the parser must start at one end of the word or the other, identifying possible allomorphs. However, rather than first generating every possible combination of allomorphs (including zero-marking allomorphs) that could comprise the word, and then eliminating the bad parses, it is vastly more efficient to stop processing a possible branch of options as soon as it can be identified as a false trail. For example, if the English word *ergonomically* is being parsed, and the parser has started at the left trying to recognize a morpheme, it may discover that the initial */er/* exists in the allomorph dictionary as the 'agentive' suffix that attaches to a verb and produces a noun. Rather that continuing to hypothesize what other morphemes might fit with this agentive one to make up the word, we want to recognize as soon as possible that this branch is a dead end. Since category mapping requires the agentive /*er*/ morpheme to have a verb to its left, it cannot be the morpheme we have found word-initially, and we can refrain from wasting any further time on this possibility, or on any based upon it. Category mapping is based on each root morpheme being assigned a specific category (such as **vi** *'intransitive verb root'*) and each affix being assigned one or more category mappings (such as **vi/V** mapping *'from intransitive verb root* **vi** *to complete verb* **V***'*). As the parser works it way through a word from one end or the other, it can be instructed to reject hypotheses where the *from category* of the currently hypothesized morpheme does not match the *to category* of the adjacent morpheme (or vice versa).

Finally, the purpose of the 'jitterbug scanner' starts to become clearer. If, in a language, categories first map rightward from the root through the suffixes and then leftward through prefixes, or vice versa, such category mapping can be handled by a left-to-right or right-to-left scan. However, if the order in which the categories are mapped alternates (as can be the case in Kiranti languages), this requires a more-involved level of processing, hence the jitterbug scanner. Watters' programmer developed such a parser, and this

strategy (among others they implemented) were found useful for our own Kiranti parsing, too.

The only other machine translation effort in Nepal that I am aware of is that of Warren Glover beginning in 1991. He successfully adapted text from the Western Gurung (Kaski District) New Testament that had been published in 1982, to the Eastern Gurung dialect (Gorkha District). The adapted books were published in 1994. He attributes the feasibility of the project to a large extent to the fact that he controlled both the source and target dialects (W. Glover, p.c. 2006).

3. OVERVIEW OF KIRANTI

3.1. Rationale for automated adaptation

According to the *Ethnologue* (Gordon 2005), there are over 100 indigenous languages spoken in Nepal. Nepali (of some variety) is the mother tongue of only about half the population. Many of the other languages are threatened or endangered, and three are already extinct. Until the 1990 revolution, the government perceived minority languages to be divisive to national unity. Nepali was the only medium of primary instruction. Since that time, there has been official recognition of the value of preserving minority languages. Many language communities themselves are looking for ways to foster language development, particularly in developing literacy in the mother tongue. Many children of minority language communities do not learn Nepali until they start school, which is taught in Nepali. It is difficult for such children to learn to read when not only are the words to be read incomprehensible, but so is the instruction itself. Consequently, they fall behind their Nepali-speaking peers. Increasingly, various language communities are working on producing transitional literacy materials, enabling children to first acquire reading skills in their own language. These skills are then easily transferred to literacy in Nepali as their Nepali language-learning catches up. However, for most minority languages, there is a crucial absence of the wide literature base that is needed next: post-literacy readers, health booklets, agricultural booklets, newspapers, textbooks, etc. Resources simply do not exist to adapt the necessary volume of existing materials into minority languages.

This is where computer-aided adaptation holds a great deal of potential. Once an adaptation tool is available, the language community can gain access to a wealth of other information and materials.

3.2. Factors Making Machine Translation More Feasible

In the last several decades, vast quantities of time and money have been poured into the quest for machine translation, and yet the results still tend to be disappointing (or amusing). What makes us imagine that we might succeed in a low-budget, resource-scarce situation as we have with Kiranti?

First, our goal is restricted: We aim only to produce *a draft that will be comprehensible* to a literate speaker of the target language. Thus, we are actually aiming only for machine-*assisted* translation. A human translator must still edit the draft in order to obtain naturalness, and even sometimes to make a selection between ambiguous alternatives. This is not as lofty of a goal as seeking a translation in which ambiguities have already been eliminated, and that needs no further editing for naturalness.

Second, with any rule-based transfer approach, the most obvious obstacle to adaptation occurs when *the source language and the target language are too different from each other*. Attempts at transfer between unrelated languages have generally produced the lowest quality of output. At the other end of the spectrum, adaptation between very similar dialects has successfully produced high-quality results. By confining this study to the adaptation within the Kiranti cluster of languages, we hope to find ourselves on the sufficiently-similar end of the spectrum.

Third, this study was based on the hypothesis that transfer can surmount structural barriers by partially encoding linguistic function. This concept will be developed further in Section 6, *The Function-Oriented Approach*. However, this strategy offers hope that even if the languages are not quite so closely related, meaningful adaptation can still succeed.

Thus, the question becomes: Given our restricted goal, and using our function-sensitive strategy, are these Kiranti languages sufficiently closely related to make automated adaptation feasible?

3.3. The Multi-Target Strategy

The investment in setting up an adaptation system may be more worthwhile if it can be carried over to other Kiranti target languages. Thus, our system has been designed from the start with this multi-target strategy in mind. The implications are developed further in Section 6.2 on the Intermediate Form, and in Section 8 on the disambiguation cycle.

3.4. Kiranti Languages

Kiranti is a cluster of languages centered in eastern Nepal. This cluster is often referred to as *East Himalayish*. Bradley (2002) places Kiranti within a larger grouping named *Himalayan* (corresponding to van Driem's 2001 *Mahakiranti "Greater Kiranti"*), which includes relatives such as Newar (of central Nepal) and Kham (of western Nepal). The Himalayan/Mahakiranti grouping itself is classified as a sub-grouping of *Bodic*, which also includes Tibetan as a distant co-descendant of Proto Tibeto-Burman.

Depending on who is doing the classification, there are between thirty and forty Kiranti languages. Weidert (1985) sketches a rough division of the major Kiranti languages as depicted in Figure 3 below:

Figure 3 Rough Linguistic Proximity of Kiranti Languages

One of the most immediately striking features of Kiranti languages is the complexity of the verbal morphology. Transitive verbs are typically marked for agreement with both the agent and patient participants, with a four-way

person split and a three-way number split, which we shall examine in Section 5.2. Several other kinds of affixes which we shall examine further complicate the verb.

Another remarkable feature of Kiranti languages is the encoding of vertical space—higher, lower, or same level—in the domain of deixis, adverbs, and case-marking. In no other grammar is vertical encoding so pervasive (Ebert 1994). The Kiranti peoples live in what is arguably the world's steepest inhabitable terrain. The Gangetic plain rushes upwards to the Himalayan peaks, a gain of up to 29,000 feet in just the hundred-mile width of Nepal. Clinging to the steep hillsides between these extremes are Kiranti villages. Whether Kirantis are going to their terraced fields, the neighbor's house, or another village, the most significant logistic factor is typically the vertical component. Obviously, in such a world, people care about the details of up/down relationships, and so it is not surprising that their languages do too.

3.5. Selection of Languages

3.5.1. Source Language Selection

One of the criteria in the selection of a good source language is that *there must be a wide literature base available in the language.* Of all the Kiranti languages, Limbu best meets this requirement. In the neighboring Indian state of Sikkim, Limbu is taught as a subject for all classes from 1 through 12; in Nepal, the government's curriculum development unit has completed Limbu instructional material up through class 4; several Limbu non-government organizations have produced literacy materials and are now producing various kinds of literature, etc.

The other major criterion of a good source language is that *adaptation should be in the direction of more complex/specified to simpler/less specified.* That is, if structural complexities exist only in one of the two languages, an adaptation process can flatten out complexity more easily than it can produce

it where it did not previously exist. For a simple lexical example, consider an ambiguous term in English, *brother*. Nepali, as most languages of the region (regardless of language family), has no direct equivalent. Nepali has a word *dāju*, meaning 'older brother', and a word *bhāi*, meaning 'younger brother.' An adaptation process from Nepali to English would have no problem translating both *dāju* and *bhāi* as *brother*, but an adaptation process from English to Nepali would face a substantial obstacle whether *brother* should be translated as *dāju* or as *bhāi*. Translating the even less specified term *sibling* to Nepali would be even more difficult. Similarly on a morphological level, if one language encodes information that is not specified by the grammar of the other, it will be difficult to automatically adapt from the less specified to the more specified.

On this criterion, Limbu again makes an ideal source for adaptation, as it is possibly the most complex and specified Kiranti language, certainly of eastern Kiranti languages.

3.5.2. Target Language Selection

Once a source language had been identified, the over-arching question this study needed to answer was, *How far can automated adaptation from Limbu reach?* The optimistic version of this question was, *Can Limbu be adapted to all of Kiranti?* To answer that question, it is necessary first to identify the Kiranti language that is:

a) among the most divergent from Limbu, and is
b) the most structurally complex.

Thulung may well be that language. Ebert's (1994) comparative grammar highlights a large number of differences between Thulung as a NW Kiranti language and Limbu as a SE Kiranti language. She characterizes the SE Kiranti languages as mainly agglutinative, while Thulung involves much in the way of stem variation and portmanteau forms. If it could be established

that automated adaptation from Limbu to Thulung is possible, then it should also be feasible to do so for every other Kiranti language.

However, a prior, more basic question exists: *Is automated adaptation feasible from Limbu at all?* Are the differences between Kiranti languages too great to permit automated adaptation? The best language for assessment seems to be Yamphu. It is not only among the varieties more closely related to Limbu, but an assessment is made feasible by the existence of the excellent descriptive grammar, *Yamphu: Grammar Texts and Lexicon* (Rutgers 1998).

Thus, the focus of my study has been on adaptation from Limbu to Yamphu.

3.5.3. Nature of Collaboration

This feasibility study or pilot project has been a collaborative effort, performed in part as research projects under the Centre for Nepal and Asian Studies (CNAS) at Tribhuvan University in Kirtipur, Nepal. Jeffrey Webster is a Limbu scholar formerly at CNAS whose focus was on Limbu analysis. Marius Doornenbal is a computational linguist whose wife was a doctor in rural Nepal. He initially tackled an assessment of transfer into Thulung, and has more recently been investigating the analysis of Bantawa, another Kiranti language. My own particular efforts were directed at transfer into Yamphu, establishment of a conceptual framework necessary for a multi-language target approach (such as the disambiguation/feedback cycle), and establishment of an "intermediate form" for Kiranti transfer. Further developments to the intermediate form were necessarily negotiated between the three of us (e.g. how to represent Limbu's multi-functional *<-aŋ>* morpheme).

3.6. Language background information

3.6.1. Limbu

Language Name

The language is typically referred to as "Limbu", although the indigenous term is *yakthuŋba pān* or *yakthuŋ pān*. Limbus may refer to themselves as *yakthuŋ* or *yakthumba*. The region in which they live is known as *pallo-kirānt,* 'Far Kirant,' or *limbuwān,* 'Land of the Limbus' (Grimes 1996).

Speakers

There are over a quarter of a million speakers of Limbu dialects. Over 90% percent of Limbus live in Nepal, in the Eastern hills. There is also a significant Limbu population over the border in India, particularly in the state of Sikkim. Limbus are traditionally farmers, growing corn, millet and rice, and raising livestock, including pigs. Literacy is about 40%. About 48% of Limbu men have completed 5 years of school, while only 6% of Limbu women have done the same. Educated individuals generally have a good proficiency in Nepali, while the less educated typically have no more than basic proficiency (Grimes 1996).

Linguistic Research

Limbu has a long literary tradition, with an orthography that originated in the early 18th century. Limbu data was collected in the 19th century and published in Grierson's *Linguistic Survey of India* in 1909. The first major linguistic work devoted to Limbu was H.W.R. Senior's *A Vocabulary of the Limbu Language of Eastern Nepal*, published in 1908. Neither of these early works transcribed the forms adequately. Since the 1960s there have been a number of papers written on various aspects of Limbu grammar, notably by authors including R. K. Sprigg, Boyd Michailovsky, A. Weidert, and George van Driem. The first thorough attempt at describing the grammar and

lexicon of Limbu was *Concise Limbu Grammar and Dictionary* by A. Weidert and B. Subba in 1985. Van Driem's (1987) grammar remains the definitive reference work, later updated by a paper entitled, *The Limbu verb revisited* (van Driem 1999). It is based on the *Phedāppe* dialect. Webster 2000 (and hence this project) is based instead on the *Pãcthare* dialect.

3.6.2. Yamphu

Very little linguistic information on Yamphu was available prior to Rutgers' Yamphu grammar. (Unless indicated otherwise, all data on Yamphu here is from Rutgers' 1998 grammar.)

Language Name

Yamphu Rai is the typical Nepali term used by this community to refer to themselves. The term "*Rai*" is often used as a synonym for "*Kiranti*," though it is more of a geographic term (Ebert 1994), and excludes Limbu. In their own language, they refer to themselves as *Yakkhaba,* and to their language as *Yakkhaba khap*.

Speakers

There are approximately 2,000 speakers of Yamphu. They live in the Arun valley in the middle hill country of eastern Nepal. Their nearest neighbors are the Mewahang Rai, the Lohorung Rai, and the Yakkhas. The Taplejung dialect area of Limbu can be reached over a snowy 17,000-foot pass to the east. Interestingly, according to a Limbu contact of Webster, some Limbu areas forbid intermarriage with any non-Limbu community *except for Yamphu*.

4. ISSUES OF WORD FREQUENCY

The simplest approach to related-language adaptation is the use of a Translation Memory (TM) system. These are used commercially in many situations with or without initial parallel corpora. The system remembers how you last translated a particular word, phrase, or sentence, and offers it again for reuse. These are advertised as "any language" systems. No linguistic rules are required, so it can be performed by a translator without requiring the involvement of computational linguists.

One such adaptation program that we considered for the Limbu-to-Yamphu transfer was a TM program produced by SIL International called *Adapt-It*. This program has been used successfully in many language pairs around the world. The impression I received, however, as I looked into this strategy, was that Limbu might not be a good candidate for transfer. The verbal morphology in particular is complex. It seemed that this could result in a longer list of individual word-forms that ultimately require manual translation. Since this is rather impressionistic, I decided to assess what empirical data might indicate regarding the relative frequency of words in Limbu.

To make a meaningful cross-linguistic comparison of word frequency, one can readily recognize the problem of comparing corpora of differing genres: A corpus of classic literature will have a much wider vocabulary than a corpus comprised only of athletic training exercises. Even if the subject matter matches closely, merely obtaining a fixed 100,000-word corpus in each language may still result in some skewing, as the same semantic content will result in corpora of different sizes for different languages. To account for this, each of the corpora should be comprised of semantically equivalent content, regardless of how many words each corpus is broken into.

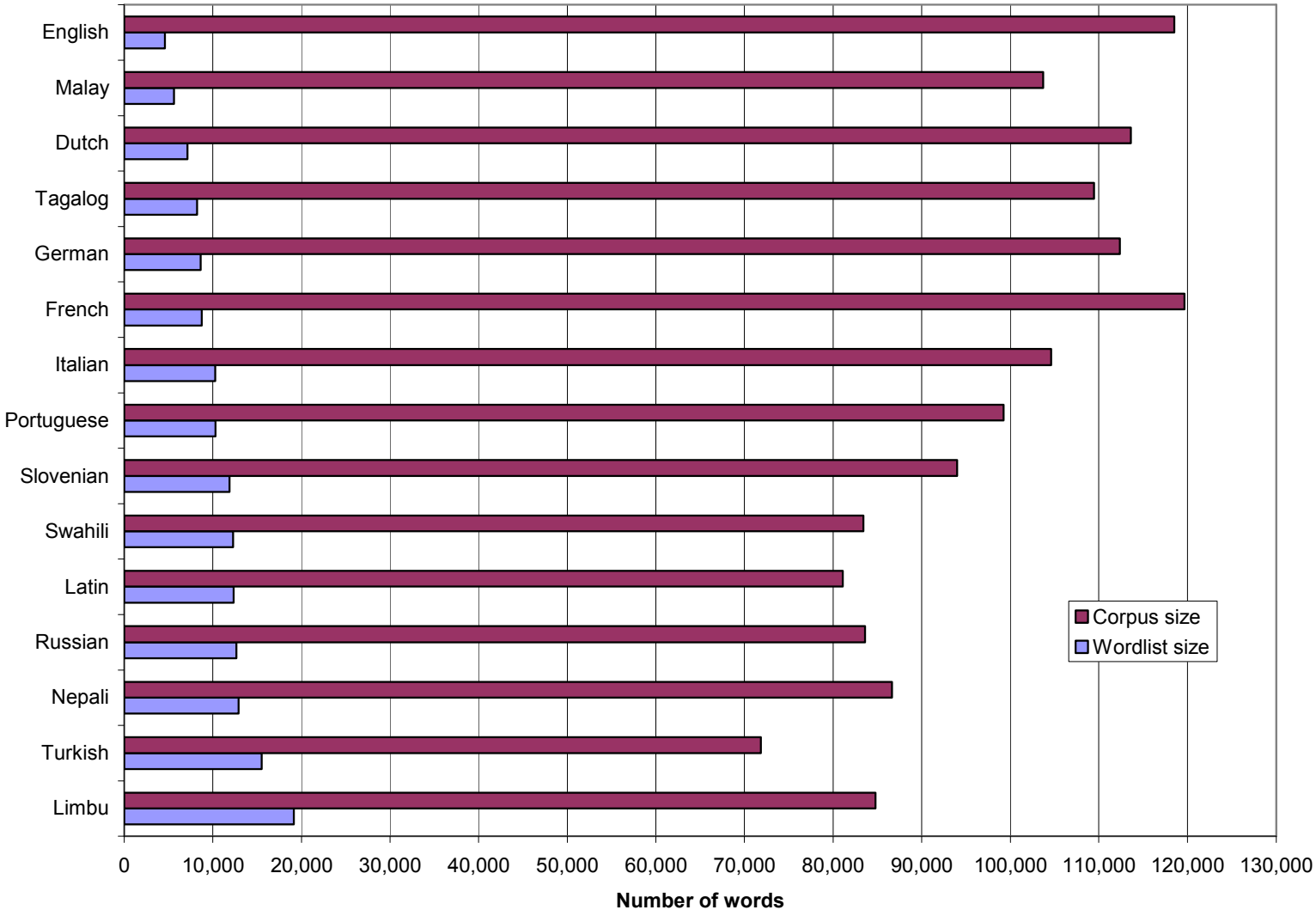
4.1. Overview of the Corpora

I was able to obtain and prepare semantically-parallel corpora for fifteen languages, including Limbu. The corpus was comprised of eighteen New Testament books, a corpus of around 100,000 words, depending on the language. For each language, I generated a wordlist of unique word-forms. The results are listed numerically in Figure 4, and charted in Figure 5.

Figure 4 Comparative sizes of the parallel corpora

Language	Wordlist size	Corpus size
English	4,599	118,498
Malay	5,604	103,714
Dutch	7,133	113,605
Tagalog	8,216	109,453
German	8,613	112,352
French	8,769	119,648
Italian	10,293	104,608
Portuguese	10,326	99,233
Slovenian	11,856	94,016
Swahili	12,291	83,411
Latin	12,336	81,081
Russian	12,664	83,610
Nepali	12,894	86,625
Turkish	15,510	71,859
Limbu	19,145	84,780

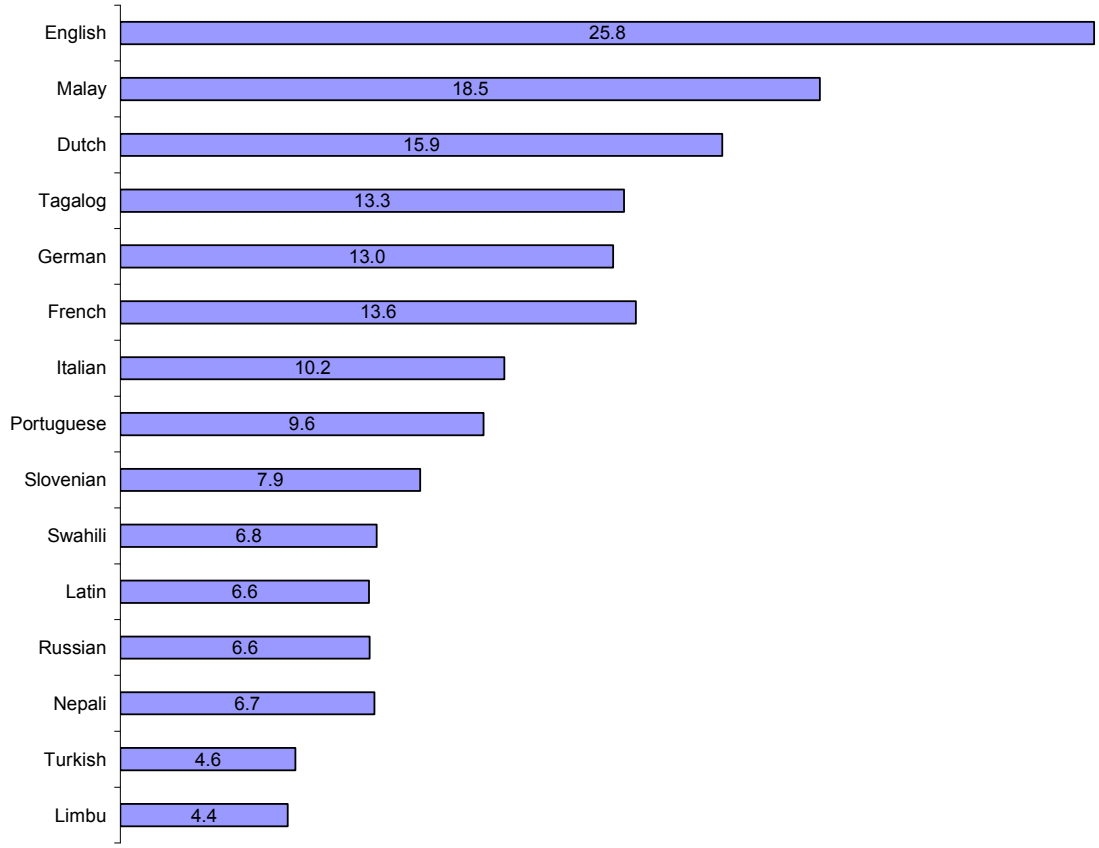
Figure 5 Visual comparison of the sizes of the parallel corpora



The dramatic difference within a language between the wordlist size and the corpus size is not the point here. In theory, had we used a smaller corpus, it would have outstripped the wordlist less dramatically, because the bigger the corpus, the closer we get toward an exhaustive wordlist. Obviously, if you could graph this for a corpus that was approaching an infinite size, there would come a point after which the wordlist would not grow much further, per the law of diminishing returns.

The comparison across languages, however, reveals more significant contrasts. Note the tendency of corpora comprised of fewer words to have longer wordlists (e.g. Limbu). Conversely, the corpora comprised of more words have shorter wordlists (e.g. English). Since the average frequency that a word appears in the corpus is the corpus size divided by the wordlist size, these two factors conspire together to emphasize the cross-language contrasts, as depicted in Figure 6 below.

Figure 6 Average frequency of word use in each corpus



Since the average Limbu word appears in the corpus 4.4 times, if the *Adapt-It* software is used here, the average word manually translated will be available for reuse another 3.4 times. In contrast, from a French corpus, the average word would be available for reuse for another 12.6 times. Clearly, this is a less productive tool for Limbu than it would be for any of the other languages in our sample.

4.2. High-Frequency Words

The use of an average word frequency, however, is perhaps misleading. In a corpus of any language, a relatively small percentage of word-forms comprise a disproportionately large portion of the corpus. Grammatical function words tend to top this list. For example, in the English corpus, the fifteen most frequently used words are as shown in Figure 7 below.

Figure 7 Most frequently used words in the English corpus

Times used	Word
7,444	the
4,745	and
4,010	to
2,814	of
2,758	you
1,863	in
1,689	he
1,513	that
1,484	a
1,383	they
1,348	I
1,281	him
1,267	will
1,177	is
1,124	who

These words comprise 0.33% of the English wordlist, and yet they represent over 30% of all words in the English corpus. Strikingly similarly in Limbu, the most frequent 0.33% of the wordlist represents about 30% of all words in the Limbu corpus. In Limbu, however, 60 different word-forms make up that 0.33% (as shown in Figure 8), four times more than the English words. Furthermore, while this set of Limbu does include grammatical words, such as conjunctions and pronouns, mixed among them are other lexical items that are either generally common (e.g. the noun *niŋwaʔ "mind",* or inflected forms of the verb root *mɛtt "say"*), or else topical in this particular corpus (e.g. *yɛsu "Jesus"*).

Figure 8 Most frequently used words in the Limbu corpus

Word Count	Word	Gloss
2,202	hɛkkʏaŋ	*and_then*
1,336	nu	*and*
1,330	kʰɛn	*that*
1,056	iŋgaʔ	*prn1s*
1,045	kʰunɛʔ	*prn3s*
1,025	kərə	*but*
863	kən	*this*
855	pʰaʔaŋ	*speech.SUB*
827	kʰiniʔ	*prn2p*
711	yɛsurɛ	*Jesus-ERG*
653	kʰuniʔ	*prn3p*
539	niŋwaʔ	*mind*
524	kʰɛnhaʔ	*that-PL*
504	tʰeaŋbʰɛllɛ	*because*
498	kʰɛllɛ	*that-ERG*
490	kak	*everyone*
490	kʰɛnɛʔ	*prn2s*
467	niŋwaʔpʰumaŋ-ŋillɛ	*God-DEF-ERG*
437	kʰɛnhaʔrɛ	*that-PL-ERG*
430	kʰɛpmo	*there.DIS*
314	be	*Qtag*
293	ɔkkʰe	*like.this*
293	yərik	*much*
288	allə	*now*
285	mənɛhaʔ	*man-PL*
284	aniʔ	*prnip*
273	cogullɛ	*do-3s-TEMP*
259	mɛttusi	*say-3p-NS*
255	waʔro	*be.exis-NP-ASS*
240	paːnnin	*utterance-DEF*

Word Count	Word	Gloss
238	tʰeaŋ	*why*
236	iksadiŋ	*earth.1*
234	yammo	*again*
232	hɛkkellɛ	*like.that-DEF-ERG*
228	yɛsu	*Jesus*
226	mɛttu	*say-3p*
225	kusiŋ	*like*
222	yapmi	*man*
217	paːnhaʔ	*utterance-PL*
210	mənɛhaʔrɛ	*man-PL-ERG*
197	tʰarik	*until*
189	nogəp	*answer*
188	weʔ	*other*
185	wəyɛro	*be.exis-PT-ASS*
184	nəsaːn	*faith/belief*
184	sese	*holy*
182	tagɛra	*almighty*
177	paːn	*utterance*
170	ləʔrik	*saying*
168	hɛkke	*like.that*
167	pʰaʔgrə	*if*
166	bi	*Q*
166	yɛsun	*Jesus-DEF*
166	abaŋe	*own*
163	anigɛʔ	*prn1p*
163	tʰik	*one*
159	coːkma	*do-INF / be.desc-INF*
157	təgi	*before*
155	waʔ	*be.exis-NP*
155	wəyɛ	*be.exis-PT*

That some lexical items appear multiple times in this list, with different case or other marking, begins to confirm our impression of the increased workload that a translation memory system would be up against.

4.3. Low-Frequency Words

An examination of the low-frequency words, however, seals the case. In a wordlist based on the corpus of any language, a disproportionately large portion of the wordlist is used relatively seldom. Indeed, a substantial percentage of word-forms appear only once. Such words are referred to as *hapax legomena*, Greek for 'read only once' (Manning and Schütze 1999).

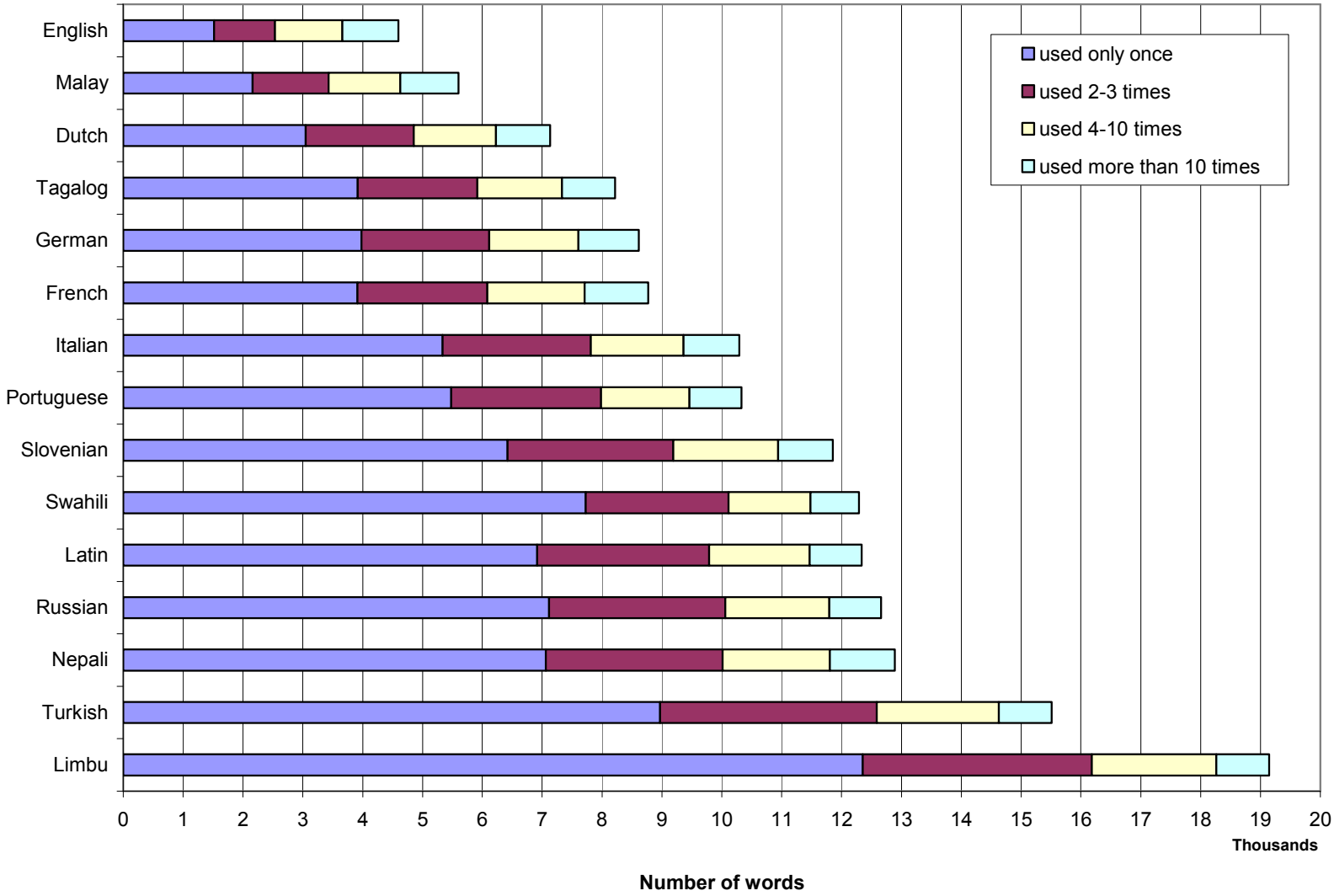
Tallying the frequency of each word-form in each of the corpora, the following results emerge:

Figure 9 Frequency of use of wordlist items

	Number of times used			
	1	**2-3**	**4-10**	**>10**
English	1,517	1,019	1,125	938
Malay	2,162	1,271	1,198	973
Dutch	3,049	1,804	1,377	903
Tagalog	3,918	2,001	1,411	886
German	3,982	2,135	1,485	1,011
French	3,911	2,174	1,624	1,060
Italian	5,336	2,477	1,545	935
Portuguese	5,483	2,502	1,475	866
Slovenian	6,421	2,769	1,751	915
Swahili	7,723	2,389	1,368	811
Latin	6,919	2,870	1,680	867
Russian	7,116	2,941	1,739	868
Nepali	7,059	2,955	1,795	1,085
Turkish	8,968	3,622	2,036	884
Limbu	12,354	3,827	2,079	885

Figure 10 displays this data graphically:

Figure 10 Frequency of use of wordlist items



Here we get a clear picture of the scale of task a TM program faces in working from a language like Limbu, as compared to the other languages shown. Recall that the *hapax legomena* items have no potential for reuse. Limbu's single-use items outnumber the entire wordlist of most of these other languages, and is more than double the size of the entire Malay wordlist.

4.4. Conclusions Regarding Word Frequency

The sheer size of the Limbu wordlist indicates that with any Translation Memory system, a relatively large number of words will need to be translated manually. The volume of *hapax legomena* items indicates that a high number of words thus translated will probably not be encountered again. Thus, a translation memory system can be expected to be less fruitful in a language like Limbu than in the other languages listed above. (Moreover, it should be noted that this approach requires of the user a much higher level of proficiency in the source language; the user must be capable of performing the translation *without* the tool in order for it to work.)

We had the impression that the *Adapt-It* software would not be so fruitful if applied to Limbu. This empirical comparison bears that out. Rather, adaptation from Limbu requires actual morphological analysis.

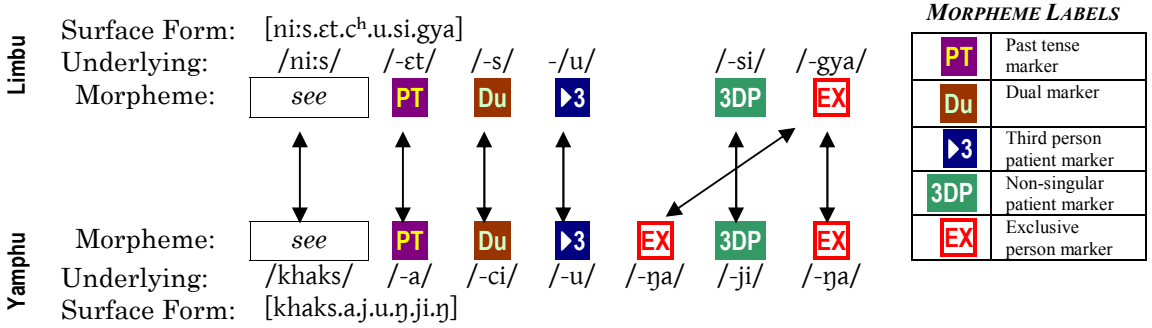
5. ISSUES OF STRUCTURAL NON-CORRESPONDENCE

5.1. An Initial Structural Comparison

What are the limits to automatic adaptation from one language to another? The typical transfer approach to adaptation focuses on structural similarities. Replace the source constituents with their target correspondents, rearrange them as necessary, and the resultant form should carry the same meaning as the source form. One of the greatest obstacles to adaptation occurs when *there is breakdown in the correspondence between source and target structures*. If a target structure does not exist in the source, how can the computer generate it?

Thus the first question in assessing the feasibility of adaptation from Limbu to Yamphu is whether these two languages are structurally similar enough. For example, compare the structure of the Limbu word *ni:sɛtcʰusigya 'We*^{de} *saw them'* with that of its Yamphu's translation, *khaksajuŋjiŋ*:

(7) 'We (dual exclusive) saw them (non-singular)'

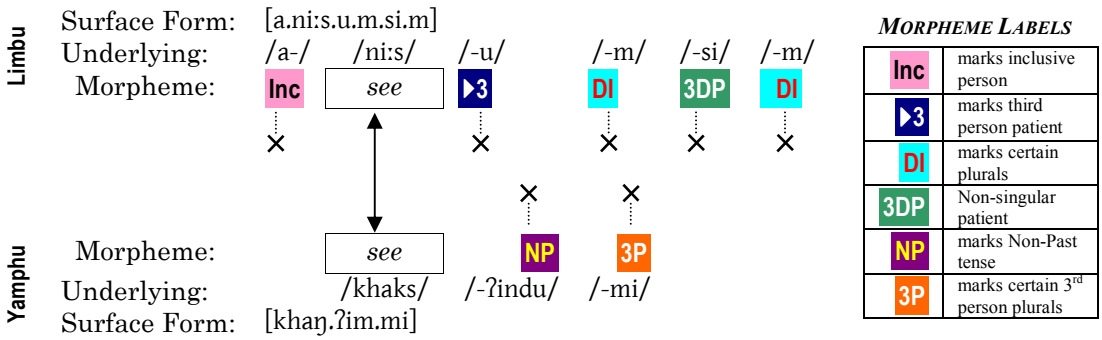


You may observe that, despite the surface dissimilarity, there is a morpheme-to-morpheme correspondence, and that the ordering of those morphemes is almost identical. Based on this example, we might hope that adaptation will be merely a matter of substituting the corresponding Yamphu morpheme and reduplicating the EX morpheme into the appropriate slot.

However, consider now the example in (8), where the meaning *"We*^{pi} *see them all"* is rendered as *ani:sumsim* in Limbu and *khaŋʔimmi* in Yamphu. The only

morpheme that has a correspondent in the structures of both languages is the verb root itself. None of the affixes carry across in either direction.

(8) 'We (plural inclusive) see them (plural)'



Before we proceed further with an investigation of this example, it is instructive to gain an overview of the workings of these affixes.

5.2. An Overview of Verbal Affixes

Each Kiranti language marks the finite verb somewhat differently, but the following categories are typically the basis for marking:

Positive or negative assertion. Positive is typically unmarked. It is a common feature among all languages of the region (including Indo-Aryan languages such as Nepali) that negation is expressed by an inflection of the verb.[4]

Tense / Modality. In Limbu, the verb paradigm makes a distinction between past and non-past tenses. In Yamphu, the paradigm distinguishes past, non-past, and perfect.

Participant Agreement: Person and/or number of the agent and/or patient. Kiranti languages grammaticalize a three-way distinction for number—singular, dual, and plural—and a four-way distinction for person: first, inclusive, second, and third.[5] Some morphemes specify grammatical

[4] That is, as opposed to using a separate negation word. (e.g. *'not'* in English)
[5] A quick glance at the Limbu agreement paradigm demonstrates that for both inclusive patients and inclusive agents, the inclusive forms pattern more like the second person forms

relations (for example, that the *agent* is a certain person/number) while other morphemes agree in some cases with the agent and in other cases with the patient. For example, in Yamphu, the same form *-c-u* (Dual Du and 3rd person patient ▶3) is used for both 2d→3s and 2s→3d. (For example, *khaŋʔitcu* is ambiguously either "*You (two) saw him*" or "*You saw them (two).*") In the former case, the dual marker *-c* Du agrees with the agent, and in the latter case, with the patient. According to Watters (2002), it is not uncommon in Kiranti languages for a disjunction to occur such that the verb agrees in person with one participant and in number with the other participant.

Tables 1 and 2 provide a bird's-eye view of how the verbal agreement affixes pattern in Limbu and Yamphu respectively. These tables reduce the affix structure of each language to the basic building blocks, enabling us to identify common patterns. The color-coding is used to aid the eye in recognizing the patterns within and between the two tables. Tense and negation markers are omitted for simplicity. Some of the Limbu affixes are prefixes, so Figure 11 indicates the position of the verb stem with V. The analysis of morpheme breaks in Limbu is from Webster (p.c. 2001). The Yamphu morpheme breaks are derived from Rutgers' (1998) analysis. The morpheme labels are my own adaptations, but it becomes clear that any labeling system is inadequate to summarize each morpheme's synchronic referential pattern.

---

than they do like the first person forms. Thus, if one were determined to use an inclusive/exclusive terminology, it might make more sense to consider inclusive/exclusive as a division of second person rather than of first person. Indeed, in diverse languages including Nama of the Khoisan family, in Yokuts of the Penutian family, and Ojibwe of the Algic family, it has been demonstrated that the morphological pattern of inclusive forms is more akin to that of second person forms. On this basis, it has been argued that in these languages, the inclusive form is more appropriately called the *second person inclusive* and the "regular" second person called the *second person exclusive*. (Harley and Ritter 2002) Here I will avoid the issue by treating *inclusive* as its own person category (1&2) distinct from 1st or 2nd.

Figure 11 Limbu Verb Affixes for Participant Agreement

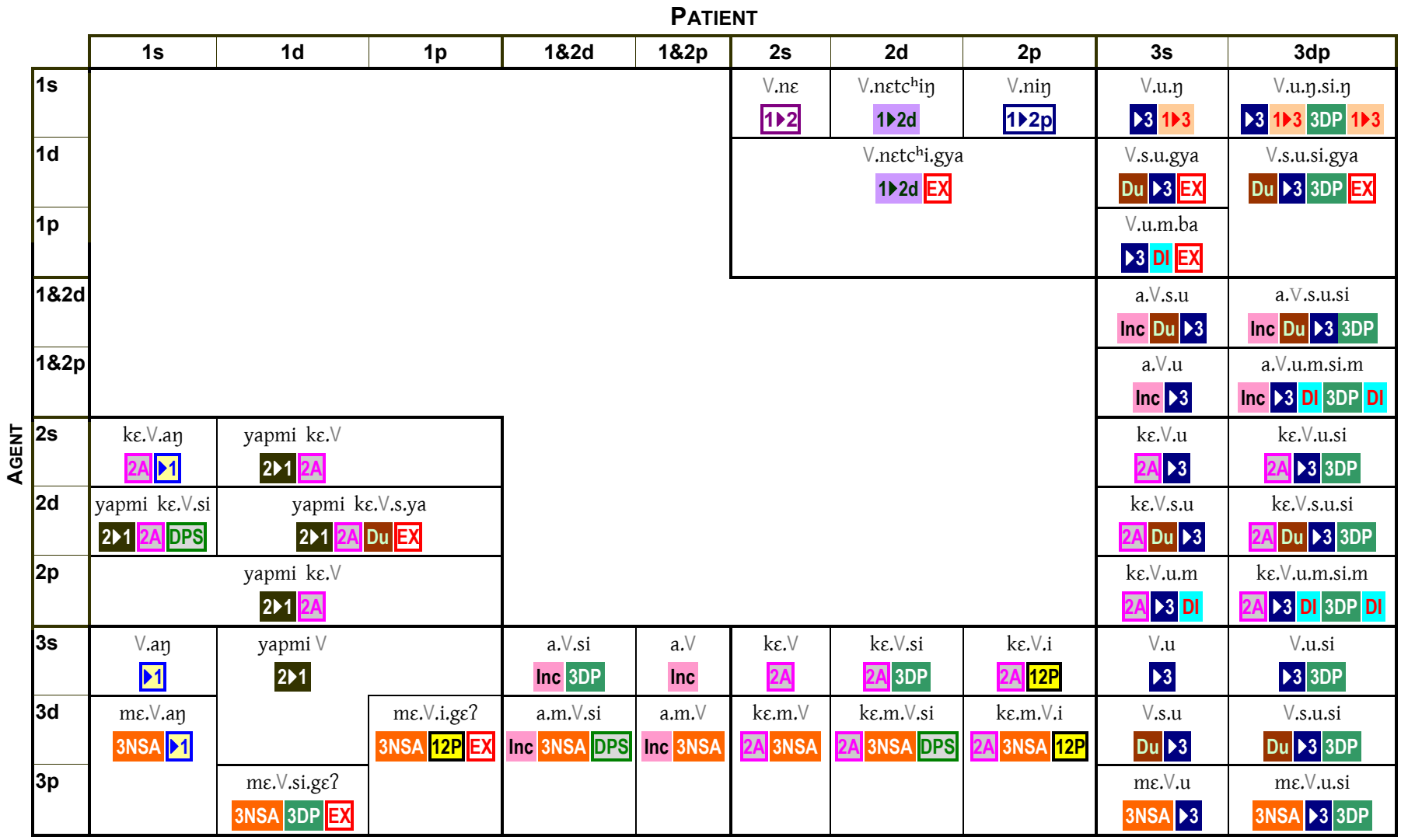
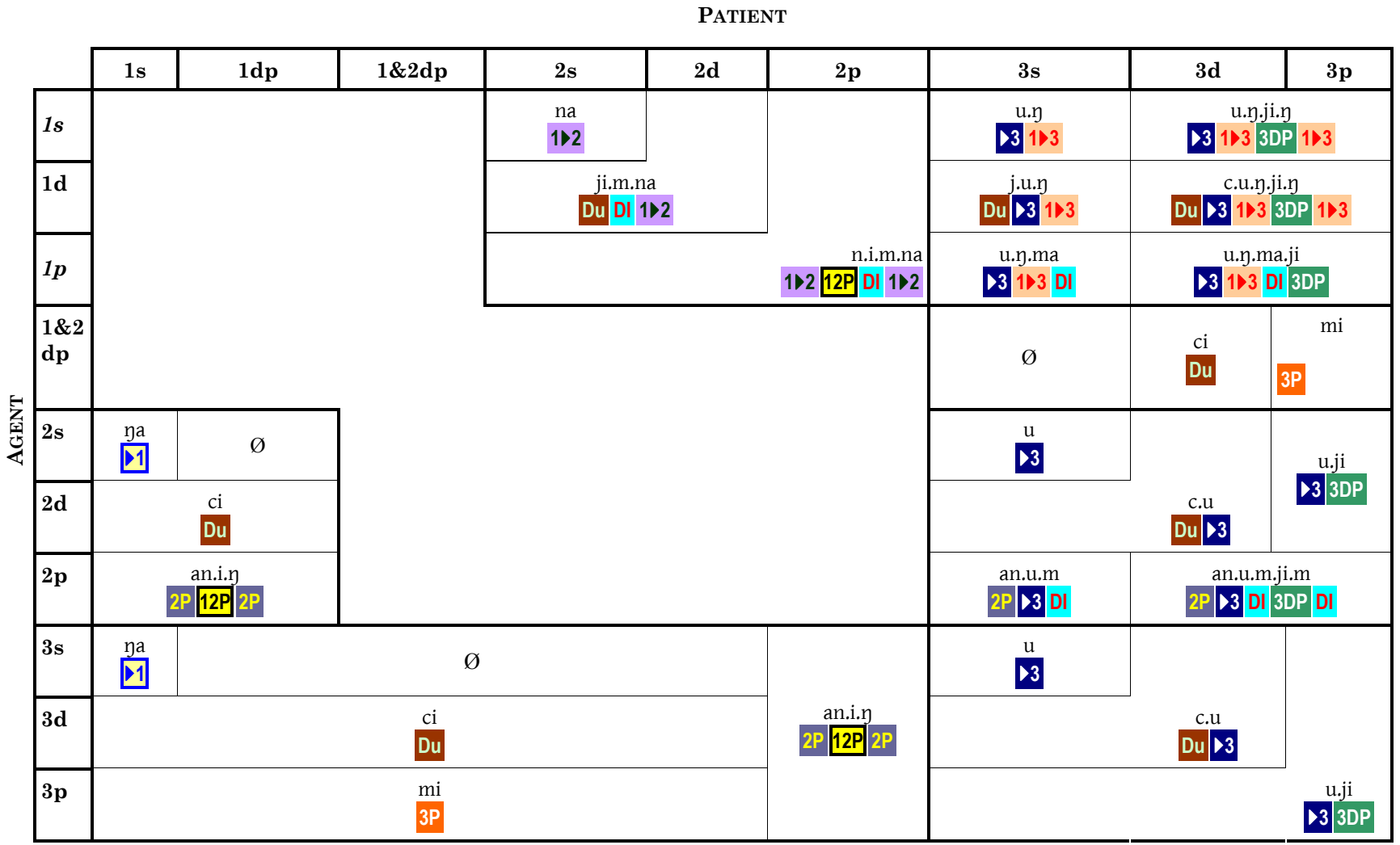


Figure 12 Yamphu Verb Suffixes for Participant Agreement

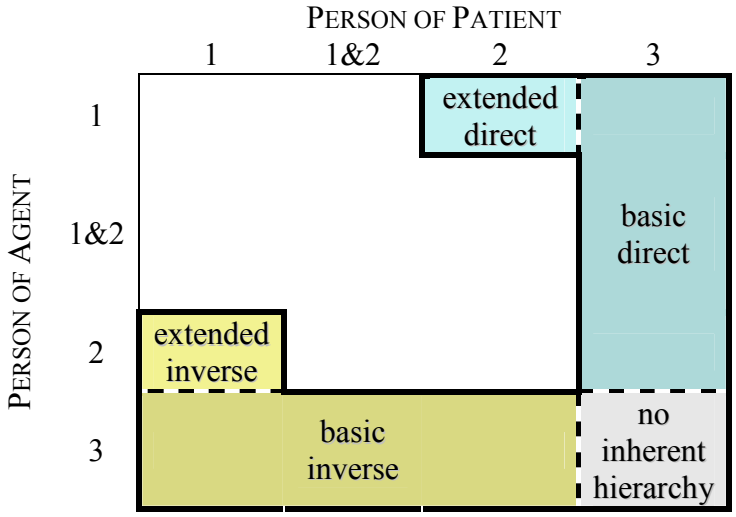


5.3. Observations on Morpheme Correspondence

5.3.1. Inverse/Direct Marking

In examining the morpheme patterns in these two tables, it may be helpful to visualize them with the following overlay. This overlay graphically depicts the inherent definitions of *inverse* and *direct* transitive configurations:

Figure 13 Inverse and Direct Configurations



Inverse/direct marking is a strategy employed in various languages of the world to mark the direction of the transitive relationship. Other person agreement markers may be present but not inherently specify whether their agreement is with the agent or the patient.

The inverse/direct marking system is a reflection of a person hierarchy. The old hierarchical pattern found in Tibeto-Burman languages is that first and second person are both ranked higher than third person: 1/2 > 3 (Watters 2002).

The first and second persons are naturally grouped here for the pragmatic reason that they are the ones involved in the speech act. The distinction in rank is thus between those who are involved in this act of speech and those who are not.

Basic inverse marking indicates that the lower-ranked participant was the agent acting upon the higher-ranked participant. Conversely, basic direct marking indicates the opposite: that a higher-ranked person was the agent acting upon the lower-ranked person.

A more refined hierarchy is also found, in which a further distinction of hierarchy is made between the speech participants, such that first person ranks above second person. The hierarchy is thus: 1 > 2 > 3. With such a hierarchy, the 2→1 configuration is also considered an inverse configuration, while the 1→2 configuration is counted as a direct configuration. This extended direct pattern can arguably be observed in Bhujel, a language that would fall in Bradley's *Central Himalayish* or Watters' *Magaranti* grouping, Kiranti's nearest relatives. An *-u/-o* suffix that is common in a variety of Tibeto-Burman languages seems to be reanalyzed from an old direct marker in the proto language (DeLancey 1981, cited in Watters 2002). In Bhujel, a seemingly conservative language, this *-u* suffix appears not only in the 1→3 and 2→3 configurations, but also in the 1→2 configuration (Watters and Regmi 2005). If this is indeed a direct marker, it implies that the refined hierarchy of 1 > 2 > 3 is present in the Mahakiranti sub-group.

Ebert (1994) notes that a number of Kiranti languages contain this marker *-u*, reanalyzed as a third person patient marker. Indeed, this stands out clearly in both Figure 11 and Figure 12 (labeled ›3) as the morpheme with the clearest referential pattern within each paradigm, and thus also with the most consistent correspondence between the two paradigms. It appears in all configurations in which the patient is third person, the only exception being

in Yamphu when the agent is the inclusive person.[6] That it even appears in the 3→3 configurations (which are neither inherently direct nor inverse) may reflect the result of reanalysis as a third person patient marker, or alternatively it may indicate that direct marking was originally applied to all configurations that were not clearly inverse. (Typically the inverse would be considered the more marked case.[7])

For the purposes of adaptation between Limbu and Yamphu, a syntactic transfer rule for this *-u* suffix (▶3) would be straightforward:

(9) Transfer rule for *-u* suffix

▶3 → Ø / Inc … __

i.e. "Delete ▶3 in a word that contains Inc."

In other words:

> "If the *-u* morpheme (▶3) is present in a source word, transfer it to the target word unless the *a-* prefix (Inc) is also present in the source word."

This environment-conditioned rule neatly and completely captures the correspondence pattern.

We will next consider the distribution of the marker that I have labeled DI. In both Limbu and Yamphu it has a phonetic form of <*-ma* ~ *-m*>. More important to adaptation than its phonetic form, however, is its referential pattern within the paradigms shown in Tables 1 and 2. For example,

---

[6] On one hand, that's a reasonable strategy for a language to adopt, because if I'm talking to you about an action in which both you and I are the agents, the normal situation is that the patient is a third person. Thus, in this case, the ▶3 marker—as a 3RD PERSON PATIENT marker—is naturally susceptible to becoming unmarked. On the other hand, considering that the original function of this *-u* morpheme was to mark the direction of transitivity, it is surprisingly unstrategic that it be dropped here, as it leaves the direction of transitivity ambiguous. The forms for "*We-two saw him / them-two / them-all*" are ambiguous with the forms for "*He / they-two / they-all saw us-two*" respectively. Perhaps that just shows how solidly reanalyzed the *-u* ▶3 marker had become.

[7] On this basis, Watters and Regmi (2005) offer an alternative explanation that although the end result is "tantamount to direct marking, its functional motivation is only the disambiguation of semantic role – an "agent identifier" (not a direction marker)."

comparing the morpheme "building blocks" of the form for 2p→3dp, we have the following:

(10) 2p→3dp

Limbu: 2A ▶3 DI 3DP DI
Yamphu: 2P ▶3 DI 3DP DI

Rutgers describes the function of this *-ma* morpheme in Yamphu as marking "non-singular number of a first or second person actant." Van Driem (1987) describes its function in Limbu as indicating "the plurality of a first or second person agent". However, there may be more to its function than that. Perhaps, just as the direct marker *-u* ▶3 was reanalyzed as a third person patient marker, the *-ma* morpheme has come to fill the role of disambiguating the direction of transitivity. It does seem to function strikingly like a [plural-tagged] direct marker. Indeed, in Limbu, this morpheme only occurs where the old direct marker *-u* ▶3 marker is also present. Limbu uses DI only in the configurations 1→3, 1&2→3, and 2→3, which are precisely the configurations that define directness in the old Tibeto-Burman hierarchy 1/2 > 3, that is, the cells in Figure 13 labeled "basic direct". Yamphu does similarly, but also extends the use the *-ma* DI morpheme to the 1→2 configuration. Thus, the true current function of the *-ma* DI morpheme may be best described as a fused *plural direct* marker, identifying the direction of transitivity as being <u>from higher person (and with plural number) to lower person</u>.

However, since Limbu and Yamphu apparently define directness slightly differently, the *-ma* DI morpheme must somehow be generated *ex nihilo* during Limbu to Yamphu transfer. For the purposes of automated adaptation, this morpheme patterns too differently in the two languages to allow for systematic transfer from one language to the other.

5.3.2. Missing Correspondents

Returning now to (8), we see that the *a-* (Inc) prefix in Limbu does not have a corresponding morpheme in Yamphu. A glance at the Limbu paradigm (Figure 11) shows that the Inc marker is used whenever either the agent or the patient is the inclusive person, and at no other time.[8] According to Watters (2002), the prefixal marking system is the older in Kiranti, and has been partially or fully supplanted by the suffixal system in Kiranti languages today. Not only does Yamphu have no prefixes, it has no agreement markers at all that correspond to the pattern of Limbu's *a-* (Inc) marker. Thus, this marker cannot be mapped into Yamphu. It would basically have to be dropped, but not before rules that refer to its presence, such as (9), have already applied.

Likewise, in the other direction, also illustrated in (8), Limbu has no morpheme corresponding to Yamphu's NP NON-PAST TENSE marker, as in Limbu this tense in largely unmarked. The Limbu analysis would have to posit a zero morpheme that can be transferred to Yamphu. That some morphemes simply have no correspondent in the other language makes the creation of structural transformation rules problematic.

5.3.3. Inconsistently-matched Correspondents

Reflected in (10) is that the referential pattern of the 3DP marker—with phonological forms of <*-si*> and <*-ji*> in Limbu and Yamphu respectively—is strikingly similar between the two languages. Thus, a clear historical

[8] van Driem (1987) labels this *a-* prefix as a FIRST PERSON marker because, in the dialect of Limbu described in his grammar, it patterns slightly differently. –Specifically, there it also appears in place of *yapmi,* that is, in the 2→1 configuration. However, he acknowledges encountering dialectal variation such that this morpheme patterns as described here. He suggests that in this dialect, the *a-* prefix has been reanalyzed from a FIRST PERSON marker to an INCLUSIVE marker. Michailovsky (1989, cited in van Driem 1999) takes this a step further and proposes to analyze this as an INCLUSIVE marker in all dialects, but van Driem does not concur, due to the use of this morpheme in first person contexts elsewhere, particularly the supine (a.k.a. "infinitive of purpose").

relationship exists, as is the case with *-ma* DI. However, as (8) demonstrates, there are cases in which the correspondence breaks down. Encapsulating these exceptions in rules presents a significant obstacle.

The final morpheme from (8) to be addressed is 3P *<-mi>*. In this case, the correspondence is fairly weak, but perhaps Limbu's 3NSA marker *<mɛ->* could be argued to show a similar pattern of usage, despite the fact that 3P is a suffix and 3NSA is a prefix.

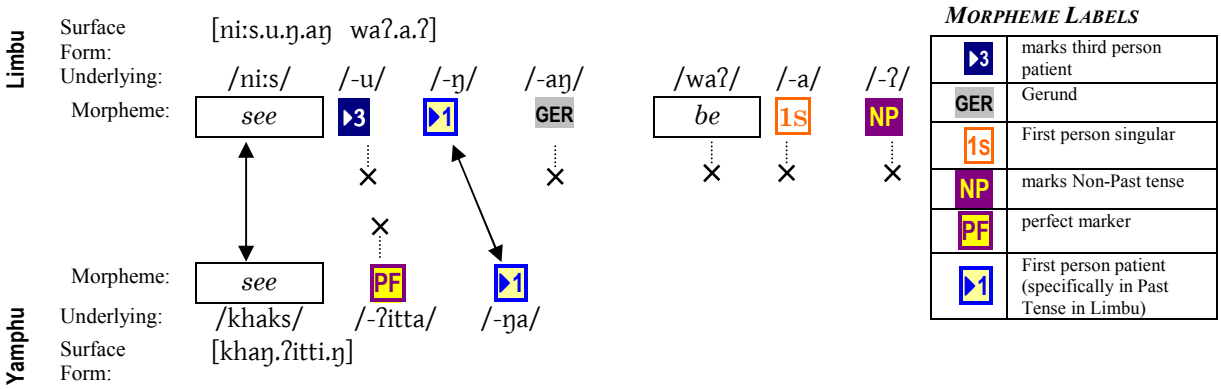
5.3.4. Conclusion

A great deal more time could be spent examining correspondences between Limbu and Yamphu morphemes. The patterns are intriguing. However, for the purposes of automated adaptation, these patterns are too complex and interwoven for systematic rearrangement. Perhaps the most important observation to make is that few of these morphemes, in themselves, provide significant actant referencing. It is the *combination/pattern of morphemes*, not individual morphemes, that convey meaning. Patterns that were once transparent now function by convention, largely the same as if the origins had instead been arbitrary. Participant agreement marking may be comprised of suffixes, prefixes, and independent words. For getting at meaning, however, the particular combinations of those markers may as well be treated as portmanteaux that may be discontinuous.

5.4. Syntactic Correspondence

Morpheme non-correspondence is not the only structural obstacle to adaptation. Lack of syntactic similarity is a further obstacle. Consider, for example, the construction of the present perfect:

(11) 'He has seen me'



In Limbu, the present perfect is a periphrastic construction comprised of a main verb plus the copula as an auxiliary with further agreement marking. In Yamphu, the perfect is marked on the verb with a different marker (PF) in the slot otherwise occupied by the PT or NP tense markers.

The structures used by Limbu and Yamphu to encode the same linguistic function (the present perfect) are substantially different. This poses challenging obstacles to automated adaptation.

6. THE FUNCTION-ORIENTED APPROACH

6.1. Overview

The approach this study has taken to the problem of structural non-correspondence is to aim for an analysis that is *functional* rather than purely *structural*. The ultimate limit to adaptation is actually function, not structure. In (11) above, despite differences in how they go about doing so, Limbu and Yamphu both demonstrate some system by which the PRESENT PERFECT is grammaticalized. This may seem unremarkable, as many languages grammaticalize the PRESENT PERFECT in one way or another, but it is the fact that two languages care about this tense-aspect that makes its translation between them possible.

Consider that in Nepali, Limbu, and Yamphu, the MIRATIVE is grammaticalized in rather different ways. Distinct from evidentiality, mirativity is the grammatical marking of unexpected, or new/unassimilated information (DeLancey 1997). Miratives are often translated into English with an exclamatory intonation pattern, as in, *"You're here!"*, or with phrases such as *"It turns out that…"* or *"Oh, I had no idea that …"* or even the genuinely surprised *"Well, whaddya know? …"* .

In Nepali, an inflected auxiliary verb (*rəhənu 'to remain'*) is used to mark the mirative:

(12) NEPALI (Mirative in positive construction)

u	gəe	**rəhecʰə**
he	go.PASTPART	**Remain.PF.3S**

'(Well, whaddya know?) –He's gone!'

(13) NEPALI (Mirative in negative construction)

kitāb	ʈebilmā	**rəhenəcʰə**
book	table.LOC	**Remain.PF.NEG.3S**

'Oh, the book is not on the table after all! (Where is it?)'

(On the other hand, in Hindi, which is Nepali's Indo-Aryan relative, the mirative is apparently not grammaticalized, and the most similar structure

to the above would apparently lend the semantics of *'to remain'* to a habitual continuous construction, which is very different from the marking of new information.)

DeLancey (1997) demonstrates that in Sunwar (a near relative of Kiranti languages, being in the Mahakiranti grouping (Gordon 2005)), the mirative is grammaticalized as an inflected form of copula.

(14) SUNWAR

kyarša	'saî- šo	**'baa-tə**
goat	kill-NOM	**MIREXIST-3SGPAST**

'He was killing a goat (I found)'

He also mentions that in Newari (also within the Mahakiranti grouping), the same semantic role is marked in the verb inflection but not in the copula, an exception among Bodic languages.

In Limbu, the mirative is marked with an uninflected sentence-final particle *ləcə/rəcə* that van Driem calls the "deprehensative particle" (DEPR). It is clearly a borrowing from Nepali's inflected *rəhecʰə* (which is typically pronounced *rəecʰə*).

(15) LIMBU

areː	hoːǃ	kɔŋ	lɛːs.u	**rəcə**
gee	whiz	this	know.3P	**DEPR**

'So, hey! He knows it!'

(van Driem 1987:241)

Yamphu's corresponding particle *læːʔæn* is not inflected either. In fact, it can be used interchangeably with the Nepali borrowing *<recha ~ rahecha>*. Rutgers refers to it as "the particle of new awareness" (NW).

(16) YAMPHU

eː	mo	tiː.beː.tt.w.e	**læːn**.di	akko
oh	that	apply.RES.PF.›3.FCT	**NW**.EXH	that

'You've put on the [cassette recorder], I see.'

(Rutgers 1998:315)

Structurally, these are all somewhat different. Yet each of these languages grammaticalize the mirative, and as a result, a basis exists by which we can transfer the mirative's meaning between these languages.

Despite differences in structural encoding, what really determines how far adaptation can be successfully carried is the extent to which languages "care about the same kind of stuff" (D. Watters, p.c. 2000).

It has already been shown that Kiranti languages care to distinguish an INCLUSIVE PERSON category and a DUAL NUMBER category. Neither of these distinctions is made in Nepali. Moreover, Kiranti is known for how its languages encode vertical space in the grammar, information that few (if any) other languages care about so deeply. If we analyze the Limbu source in terms of linguistic functions carried by its structures, rather than simply analyze what the structures are, we should be able to substantially widen the reach of automated adaptation.

6.2. The Intermediate Form

6.2.1. Multi-Language Target

If we have the further goal of generating adaptations in multiple languages, a Limbu text that has been analyzed in terms of linguistic functions (we shall refer to this as the *Intermediate Form*) provides the ideal common source for these multiple targets. Someone who is working on transferring text from this neat, logical form into a particular Kiranti target language does not require an understanding of the intricacies of the original Limbu structures.

The *Intermediate Form* is essentially an idealized analysis of Limbu. In the Intermediate Form, ambiguities from the Limbu analysis are disambiguated to the greatest extent possible. That is, the Intermediate Form aims to make distinctions that Limbu itself does not mark, disambiguating either as a result of syntactic analysis, or by the manual disambiguation cycle described in Section 8. Moreover, per our "function-oriented" approach, in the

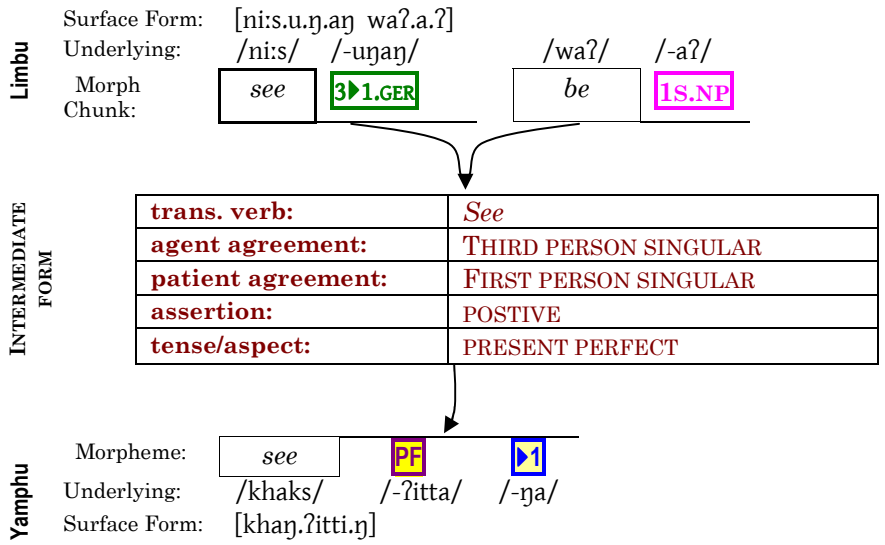
Intermediate Form, quirkiness in Limbu is ironed out into a logical representation. Part of this involves representing the semantic load or units of linguistic function as idealized pseudo-morphemes with their own neat morphology. I shall use the term *functeme* to refer specifically to these invented pseudo-morphemes. To the extent that the Intermediate Form carries a representation of semantics in place of the actual Limbu syntax, it is actually a hybrid interlingua approach to machine translation.

Thus our adaptation process includes the following basic steps:

- Analysis of Limbu constituents.
- Rearrangement of Limbu's quirky parts into *functemes* in a structure-neutral Intermediate Form.
- Rearrangement of the Intermediate Form for Yamphu structures.
- Synthesis of Yamphu surface forms.

Putting aside the implementation details, this process can be conceptualized as shown below with the present perfect example previously examined in (11):

(17) 'He has seen me'



Thus, the intermediate form provides a means of encapsulating the source meaning so that it will be accessible to multiple target languages.

6.2.2. Implications for Parsing the Limbu Source

Clumping Morphemes Together

Note that with this method, it is not necessary to analyze every individual morpheme in the Limbu source, as would be typical for an academic interlinearization. Rather, because meaning is conveyed by the combination of prefixal, suffixal, and freestanding morphemes, and because Limbu's suffixal agreement morphemes are never separated from each other, the intermediate form may be constructed based on the combinations of these morpheme clusters. In the above example, although */-uŋaŋ/* could be parsed into three separate morphemes (as in (11), where they were labeled ▶3 ▶1 GER), with the current approach, this is not necessary. (Indeed, it would complicate the analysis greatly to have to specify rules governing the environments in which a particular morpheme may appear.) Instead, the intermediate form can be generated based upon the unparsed combination labeled above as 3▶1.GER, as if it were a single portmanteau morpheme that we do not chop into smaller bits. This morpheme cluster is listed in the dictionary as a single suffix. When it is found in the pattern shown in (17), this is recognized as a 'positive 3s→1s present perfect' construction, all without the need to more finely analyze what */-uŋaŋ/* is internally comprised of.

Handling Freestanding Participant Marking

This approach also provides a standardized means of handling the freestanding word *yapmi* that appears before certain inflections of the verb. As with the other participant agreement morphemes on verbs, the referential pattern of *yapmi* is complex. It could be a kind of 2→1 marker. Following are examples of inflections with and without *yapmi*.

(18) Inflectional Minimal Pairs for yapmi on the root hip *'to hit'*

kɛhip	3s → 2s	He hits you(s)
yapmi kɛhip	2s/p → 1d/p	You hit us

kɛhipsi	3s → 2d	He hits you(d)
yapmi kɛhipsi	2d → 1s	You(d) hit me

Thus *yapmi* functions as any other participant marking morpheme in Limbu, except that it is not attached to the verb. Adjectives may even appear between *yapmi* and the verb, establishing that *yapmi* is not merely a prefix. This word can be used independently to mean "person". Whenever syntax indicates that the use is part of the agreement marking, because the intermediate form encapsulates the semantics of who is acting on whom, the *yapmi* structure is not passed through for an adaptation target language to have to handle. In all other cases, it *is* passed through to be transferred with the meaning of "person."

7. IMPLEMENTING ADAPTATION

We now turn to the issue of how this strategy can be implemented. Implementation depends largely on the particular software employed. For the reasons described in Section 4, we needed to use software that could perform morphological parsing. (While it would be possible to train statistical methods to recognize individual Limbu morphemes, a significant Yamphu corpus is necessary for the training of the translation model, even more so for the training of the Yamphu language model. Given the issues of word frequency in Limbu, it would seem that Kiranti languages would require even larger corpora than the European languages do. Therefore, statistical strategies were not employed.) A system based on analysis, transfer, and synthesis is really the only practical approach to adaptation for this situation. A number of morphological parsers have floated around in the academic community, but fewer options are available for the complete translation task.

7.1. A *Toolbox* Implementation

We began this project using *The Field Linguist's Toolbox* (or just *Toolbox* for short) by SIL International. *Toolbox* is the successor to the *Shoebox* software. In this approach, the text to be transferred begins as a text file formatted with Standard-Format-Marking (SFM).[9] The analysis stage is much like interlinearization, in that we produce a line of morpheme glosses, lined up under the source text. In the transfer stage, these glosses are rearranged according to the needs of the target language. In the synthesis stage, the target language's morphemes are substituted, and merged in accordance with phonological rules. Thus, the entire process can be visible in a single SFM

[9] SFM is an SIL format for text data. It's a fairly loose and transparent format, predating SGML/XML. Each field is simply identified by a line of text beginning with backslash character (\) followed by one or more text characters that comprise the *field code*.

text file. Here is a simplified example of how a phrase might transfer from Limbu to Yamphu:

(19) Example of basic adaptation in Toolbox
'to tell a lie in the pasture'

\lt	caramdɛnno			iŋlɛk		mɛpma			*Limbu text*	
\lm	caramdɛn	-o		iŋlɛk		mɛtt	-ma		*Limbu morphemes*	
\g1	pasture	-LOC		lie		say	-INF		*Limbu gloss*	
\p1	*n*	*-case*		*n*		*vt*	*-NOM*		*Limbu part of speech*	
\g2	pasture	-LOC				lie	-INF		*rearranged gloss*	
\p2	*n*	*-case*				*vi*	*-NOM*		*rearranged part of speech*	
\ym	caura	-peʔ				remded	-ma		*Yamphu morphemes*	
\yt	caurabeʔ					remdeʔma			*Yamphu text*	

Source: Line **lt** contains the **L**imbu source **t**ext to adapt.

Analysis: Given **lt**, a parse process produces line **lm**, which contains the **L**imbu **m**orphemes in their dictionary form. From that dictionary entry, it looks up the morpheme's unique gloss (copying it to the **g1** line) and the morpheme's part of speech (copying it to the **p1** line).

Transfer: Given **g1** and **p1** (unique gloss and part-of-speech respectively), a rearrange process adjusts the analysis for Yamphu, producing lines **g2** and **p2** respectively. In the above example, an adjustment is made at this stage to replace the Limbu structure (noun *lie* plus transitive verb *say*) with the Yamphu structure (intransitive verb *lie*). The part-of-speech category provides a means for rearrange rules to take advantage of generalizations. The value in the part-of-speech field may be more finely refined than the broad categories often conceived of as the part of speech, in order to make distinctions in verb valencies, or in the more nitty-gritty details of what it may co-occur with.

As this is a simplified example, only one rearrange process is shown here. Although one rearrange process may involve many applicable rules, in practice, multiple rearrange processes are necessary in order to manage how

the output of one rearrange process will be handled as the input of the next. Thus, not shown are lines **\g3** and **\p3**, **\g4** and **\p4**, and so on.

Synthesis: From the Yamphu dictionary, the Yamphu morpheme having the same gloss as that on the final **\g*N*** line is copied to the **\ym** line. This is, of course, the underlying form of the morpheme. Phonological and morphophonemic rules are applied when the morphemes are joined, resulting in line **\yt**, the Yamphu text. These rules may feed each other, and so the final surface form may be significantly different to the underlying form.

In synthesizing the surface form in the target language, *Toolbox* is clearly superior to *CarlaStudio*, an alternative software package for adaptation, which will be discussed in the next section. A target morpheme is entered in the lexicon in its underlying form. Phonological and morphophonemic rules are specified in a separate rule table. With *CarlaStudio*, phonological rules may be applied to roots, but for affixes, all the allomorphs must be specified, along with the phonological contexts in which they surface. Dealing with the surface forms instead of the underlying forms is linguistically less elegant.

Adaptation in *Toolbox* is an interactive process. If an ambiguity arises that cannot be solved by any already-provided rules, *Toolbox* immediately prompts the user to make a selection. Obviously, this may happen at the parse stage, when more than one parse is possible. It can also be made to happen at the synthesis stage, if the target dictionary is given more than one entry with the same gloss. For example, this can be utilized to prompt the user to select between two forms that are distinguished in Yamphu but not in Limbu.

Toolbox's disambiguation rules enable it to properly identify morphemes based on word-level rules. For example, *Toolbox* can analyze the *-s* suffix in *cats* as a PLURAL marker, while in *sings* it can be analyzed as 3RD PERSON SINGULAR agreement, because *cat* is a noun while *sing* is a verb. However, this disambiguation cannot refer to the syntax beyond the word level. For example, *Toolbox* cannot automatically analyze *knocks*, but requires the user

to first specify whether *knock* is a noun or verb in the current context (e.g. '*He knocks hard'* vs. '*He got some hard knocks*').

Toolbox's rearrangement rules are able to "see" syntax patterns and apply wider transformations accordingly. In theory, it is possible to take advantage of this to resolve ambiguities syntactically. To continue with the English example of *knocks*, rather than having two dictionary entries for *knock* (one with a part of speech of ***n*** and the other of ***vi***), it would be possible to list *knock* in the dictionary with a "part of speech" that indicated that this could be either a noun or a verb (for example: **nvi**) and to instruct *Toolbox* to assume that *-s* means PLURAL after an **nvi**. Then, a rearrange rule can identify syntactic patterns, such as **NP nvi -PLURAL** that should be rearranged into **NP nvi -3sAGR**. However, this sparing of user interaction comes at a price of complicating the dictionary and the rearrange rules.

The rearrange rules also turn out to have a rather significant short-coming: They can only "see" the current line. In *Toolbox* it is not uncommon for sentences to wrap around onto a second line, but when this happens, the rearrange processes do not see the rest of the syntactic context. Primarily for this reason, we made the decision to start our transfer and synthesis work over with the *CarlaStudio* suite of programs, also by SIL. Initially, we kept the Limbu analysis in *Toolbox*, doing as much disambiguation there as could be done with reference only to word-level rules. Any morpheme that could not be disambiguated at that level had to be passed through to *CarlaStudio* for syntactic disambiguation. For example, the *-illɛ* suffix in Limbu may mark ERGATIVE-INSTRUMENTAL (Kiranti languages do not make a formal distinction between ergative and instrumental), GENITIVE, or TEMPORAL, which are all homophonous. TEMPORAL marking is a verbal suffix, so *Toolbox* could immediately identify the instances for which the *-illɛ* suffix should be thus marked. On the other hand, where the suffix appears elsewhere, it could indicate either GENITIVE or ERGATIVE-INSTRUMENTAL, depending on the

context. Most other Kiranti languages have a genitive form that differs from that of the ergative-instrumental, and thus the Intermediate Form needs to make this distinction. Toolbox would merely tag the suffix as -ERGGI and then *CarlaStudio* could apply syntactic tests to determine whether to output this to the Intermediate Form as -ERG or -GEN. (cf. Section 9.1.1) Later on, however, even the Limbu analysis was re-implemented in *CarlaStudio*, for the sake of maintainability.

7.2. A *CarlaStudio* Implementation

CarlaStudio is actually a set of SIL programs that each handle specific parts of adaptation. CARLA is an acronym for Computer-Aided Related-Language Adaptation. The most significant components are AMPLE (**A M**orphological **P**arser for **L**inguistic **E**xploration), SENTRANS (**Sen**tence **Trans**fer) and STAMP (**S**ynthesizing after **T**ransferring **AMP**LE Analyses). A variety of other components can also be utilized, including CC (**C**onsistent **C**hanges) and any special-purpose software developed by the user for unique needs. As the CARLA programs were not written to handle our function-oriented approach or multi-target strategy, I developed some external Perl scripts for these purposes.

There are a number of significant differences between CarlaStudio and Toolbox implementations:

Whereas in *Toolbox*, each step of the adaptation process results in another line or two *within the same file*, in *CarlaStudio* each step of the adaptation process results in the generation of a modified version of the file.

Unlike *Toolbox*, *CarlaStudio* pays no attention to where line breaks fall in the source file. Rather, sentences are properly identified according to punctuation.

CarlaStudio makes an important distinction between an *analysis* file and a source/target *text* file. The text file may be an SFM file that contains the

source and target text in a specified field (as in *Toolbox*), or it may be an ordinary text file containing only the unformatted source/target text. As a minimal example, consider a source text containing only one sentence:

(20) Contents of source file

```
tʰikyɛn kʰɛllɛ kuniŋwaʔo mɔnɛhaʔ iŋlɛk mɛpmasi tʏɛ
```

An analysis file is essentially an SFM database containing a record for each occurrence of each word in the text. From the source file in (20), the analysis file may look like this:

(21) Contents of corresponding analysis file

```
\a   < adj one n day:n >
\d   tʰik-yɛn
\cat n adj=n
\u   0-0
\w   tʰikyɛn

\a   < np that > ERGGI
\d   kʰɛ-llɛ
\cat np np=np/np
\u   0-0
\w   kʰɛllɛ

\a   pos3s < n mind:n > LOC
\d   ku-niŋwaʔ-o
\cat n n/n=n=n/n
\u   0-0-0
\w   kuniŋwaʔo

\a   < n man:n > PL
\d   mɔnɛ-haʔ
\cat n n=n/n
\u   0-0
\w   mɔnɛhaʔ

\a   < n lie:n >
\d   iŋlɛk
\cat n n
\u   0
\w   iŋlɛk

\a   < vt say > infp
\d   mɛp-masi
\cat vt vt=vt/vt
\u   0-0
\w   mɛpmasi

\a   < vi come.far > 3PT
\d   tY-ɛ
\cat vi vi=vi/vi
\u   0-0
\w   tYɛ
\n   .
```

Each word in the source has a record in the analysis file. Within that record, the **\a** field contains the morpheme glosses. The root is enclosed in <angle brackets>, along with its part-of-speech category. Glosses for prefixes and suffixes appear respectively before and after the root. Further information about this word is contained in other fields. The most significant fields to notice at this point are the **\d** field, which contains the allomorphs into which the word can be decomposed, and the **\cat** field, which contains the category

mappings used in the parsing process to ensure that only the right category of affixes were appended to the word in the parse process.

Another significant difference in approach from *Toolbox* is that *CarlaStudio* handles ambiguities without user interaction. If the ambiguity cannot be resolved any other way, *CarlaStudio* retains the multiple options in a special format. For example, if *knock-s* could be analyzed as either:

<*v* **knock**> 3SG.AGR

or:

<*n* **knock**> PLURAL

CarlaStudio internally represents this word as:

%2% <*v* **knock**> 3SG.AGR % <*n* **knock**> PLURAL %

The digit 2 here indicates that the ambiguity contains two alternatives.

(Note also that *CarlaStudio* represents this with plain text. The font formatting shown here and throughout this document is provided for visual clarity, but is not an inherent part of the data itself.)

Successive processes may be able to resolve the ambiguity by referring to syntactic features. If not, the percent notation continues to mark the ambiguous alternatives in the text finally synthesized, where it can be manually disambiguated.

7.3. Implementing the Intermediate Form

7.3.1. The Role of Functemization

The Intermediate Form has been described as "an idealized analysis" of Limbu, and of course the representation of an analysis is quite different in *Toolbox* and *CarlaStudio*. In *Toolbox*, the Intermediate Form was defined as a *unique identifier for every morpheme, tagged with a part-of-speech marker.* In terms of the fields shown in (19), an analysis is a pair of gloss and part-of-

speech lines, for example **\g1** and **\p1**. If that example were expanded to show multiple rearrange processes, the earlier processes would be for tidying up the Limbu analysis in various ways, for whichever Kiranti target was the adaptation goal. The final processes would be for rearranging the analysis specifically for Yamphu's structural needs. In between, a particular pair of fields (called **\gi** for "**i**ntermediate **g**loss" and **\pi** for "**i**ntermediate **p**art of speech") comprises the Intermediate Form.

In *CarlaStudio*, the representation of the Intermediate Form changes to that of an analysis file, but as its one-word-per-record representation loses some transparency, I will introduce the structure of the Intermediate Form using the *Toolbox* representation, which is visually easier to convey.

The Intermediate Form, then, is a unique identifier for every morpheme (**\ig**), tagged with a part-of-speech marker (**\ip**). The unique identifier may be arbitrary. In the current project it has been based on the English gloss.

For example, consider the following analysis of Limbu:

(22) 'There was a man grazing cows in the cow pasture.'

\lt	lɔttʰik	mɔnɛn		pit	caramdɛnno		pit	carammi		wɔyɛ			
\lm	lɔtcʰa	mɔna	-ʔin	pit	caramdɛn	-o	pit	caram	-ʔi	wa	-ɛ		
\g1	**one**	**man**	**-DEF**	**cow**	**pasture**	**-LOC**	**cow**	**graze**	**-SIM**	**be**	**-PAST**		
\p1	***num***	***n***	***-def***	***n***	***n***	***-PPos***	***n***	***vt***	***-aux***	***vi***	***-Vchunk***[10]		

The \g1 and \p1 lines, taken together, constitute an initial attempt at the intermediate form. However, our intermediate form must go beyond merely representing morphemes, as the morphosyntactic structure is not always able to be carried across. In the above example, the structure shown in **red** would certainly be insufficient information for transfer to most Kiranti languages, as no participant agreement morphemes are present. Note also that this

[10] Anything given a "part of speech" of *Vchunk* is merely one or more verbal affixes "chunked" together, for the reasons described in Section 6.2.2 above.

structure combines with the structure shown in **blue** to create an auxiliary verb construction. We need to consider the possibility that this auxiliary verb construction might not convey the same meaning in other Kiranti languages. In our function-based approach, the intermediate form must represent the *meaning* that can be carried across.

We can do this by *functemization:* Replacing actual morphosyntactic structures with *functemes*: *pseudo-morphemes that represent semantic function*, making the intermediate form represent the source as if the source had been both systematic and explicit. Thus, the conceptualization given in (23) can be implemented as shown in (24).

(23) Conceptualization of Intermediate Form: 'He has seen me'

INTERMEDIATE FORM

trans. verb:	*see*
agent agreement:	THIRD PERSON SINGULAR
patient agreement:	FIRST PERSON SINGULAR
assertion:	POSITIVE
tense/aspect:	PRESENT PERFECT

(24) Implementation of Intermediate Form: 'He has seen me'

LIMBU

\lt	ni:suŋaŋ		waʔaʔ	
\lm	ni:s	-uŋaŋ	waʔ	-aʔ
\g1	see	-3→1.GER	be	-1s.NP
\p1	*vt*	*-Vchunk*	*vi*	*-Vchunk*

↳ Functemization ↴

INTERMEDIATE FORM

	SEE	POSITIVE	PRESENT PERFECT	THIRD	SINGULAR	FIRST	SINGULAR
\gi	**see**	**-P**	**-EF**	**-3**	**-s**	**-1**	**-s**
\pi	***vt***	***-a***	***-t***	***-pA***	***-nA***	***-pP***	***-nP***
	trans. verb	assertion	tense / aspect	person of agent	number of agent	person of patient	number of patient

(The values on the \gi and \pi lines are discussed in Section 7.3.3 below.)

In some places, the intermediate form reflects the actual morphosyntactic structure. In other places, it contains function tokens. The following example shows how the intermediate form incorporates both morphemes and functemes, side by side.

(25) 'That cowherd lied to them many times in this way.'

\t	kʰɛn	pitkɔmballɛ		yɔrik	lɛŋ	ɔkkʰelɔrik	iŋlɛk	mɛttusi						
\m	kʰɛn	pitkɔmba	-illɛ	yɔrik	lɛŋ	ɔkkʰelɔrik	iŋlɛk	mɛtt	-usi					
\gm	that	cowherd	-ERGGI	much	time	this_way	lie	say	-›3dpP					
\pm	DEM	n	-nfl	adj	adv	adv	n	vt	-Vchunk					
	↓	↓	↓	↓	↓	↓	↓	↳ *functemization* ↴						
\gi	that	cowherd	-ERGGI	much	time	this_way	lie	say	-P	-PT	-3	-s	-3	-p
\pi	DEM	n	-case	adj	adv	adv	n	vt	-a	-t	-pA	-nA	-pP	-nP

The part of the intermediate form shown in **blue** transfers directly across, while the part shown in **red** is comprised of functemes.

7.3.2. Idioms in the Intermediate Form

Compound Nouns

It should be noted that even the form that transfers directly across might not necessarily be a morpheme-to-morpheme transfer. For example, the Limbu word *pitkɔmba 'cowherd'* in the above example is actually a compound noun, comprised of two morphemes meaning *'cow'* and *'shepherd.'* There are two possibilities for the intermediate form:

(26) 'cowherd'

	(a) TREATED AS A SINGLE LEXICAL ITEM	(b) TREATED AS A COMPOUND NOUN	
\t	pitkɔmba	pitkɔmba	
\m	pitkɔmba	pit	kɔmba
\gm	cowherd	cow	shepherd
\pm	n	n	n
	↓	↓	↓
\gi	cowherd	cow	shepherd
\pi	n	n	n

A judgment call is necessary in such cases. On the one hand, it is desirable to avoid having to provide a unique identifier to every such compound, as this

requires the lexicon of the target language to provide equivalents for each. Since it is possible that the semantic equivalent in the target will be constructed from the same morphemes, such treatment of compounds as single items can bloat the lexicon unnecessarily. On the other hand, in some target languages, this compounding of nouns might not be meaningful, in which case the intermediate form should contain a single noun identifier.

Idiomatic Verbs

A similar but more complex issue arises with the plethora of idiomatic verbal constructions in Limbu. For example, there is not a verb in Limbu that in itself means *'to grieve.'* In Limbu, the notion *'he was grieved'* is expressed *luŋma sɔnc^{h}ɛ*, literally *'His liver fell (laying stretched out).'* In Limbu, the liver is the seat of emotions, and thus it recurs in a wide variety of idiomatic verbal constructions. Indeed, apparently there are similar idioms in other Kiranti languages, although it may be another body part, such as the stomach, that fills this role. Thus it may be possible to substitute only the name of the body part to render the idiom comprehensible in the target language. For example, while English speakers might wonder at the intended meaning of *'His liver fell',* they could readily grasp the meaning of *'His heart fell.'* On the other hand, some idiomatic constructions might not carry across at all, or worse, they may carry across with a meaning far from that of the source language. In such cases, the intermediate form should contain the identifier of a pseudo-verb, from which each target language can construct the meaning according to its own structure and/or idiom.

7.3.3. Structure of the Intermediate Form

Just as "conjunction" as a part of speech defines a closed class of words that can fill a particular grammatical slot, so our functemes can be conceived of as belonging to closed classes, filling particular slots in the grammar of our intermediate form. Example (27) below shows how the finite transitive verb

is constructed from seven slots in the "grammar" of the intermediate form, and the values that may appear in each of those slots.

(27) Intermediate Form of the Finite Transitive Verb

	\pi tag	Description	Class	\gi value	Description
1.	vt	trans. verb identifier	open	*unrestricted*	
2.	-a	assertion	closed	-P	positive
				-N	negative
3.	-t	tense/modality	closed	-NP	non-past
				-PT	past
				-EF	present perfect
				-AF	past perfect
4.	-pA	person of agent	closed	-1	First person
				-i	Inclusive person (non-singular only)
				-2	Second person
				-3	Third person
5.	-nA	number of agent	closed	-s	singular
				-d	dual
				-p	plural
6.	-pP	person of patient	closed	-1/i/2/3	1st/incl./2nd/3rd *(as for* -pA*)*
7.	-nP	number of patient	closed	-s/d/p	sg./dual/plural *(as for* -nA*)*

It is this technique of representing an interlingual elements in a pseudo-morphology that enables us to hybridize an interlingual strategy into a transfer-based system.

8. DISAMBIGUATION

8.1. Possible Disambiguation Points

There are three possible points at which ambiguities might be manually resolved:

Disambiguation During the Limbu Parse

During the Limbu parse, it is possible to force the user to provide immediate disambiguation. For example, the user must examine the context, and select either the past tense or the non-past tense, or select between dual and plural agreement. A significant problem here is that such distinctions may be hard for a mother tongue Limbu speaker to discern. Since the language itself does not make the distinction, the distinction is less relevant to the speaker. Thinking in these terms requires both training and greater cognitive effort.

Also, it may be possible to resolve some ambiguities syntactically after the parsing phase. It would be needless effort to do so manually in such cases.

Disambiguation On the Intermediate Form

One premise of the *Intermediate Form* is that the analyzed text ought to be able to be transferred from this form into multiple Kiranti language targets. By disambiguating the Intermediate Form once, we save ourselves the effort of re-disambiguating the same ambiguities again for each target language. By waiting until it reaches this form to do the manual disambiguation, we are able to limit the manual effort to only the ambiguities that cannot be resolved by syntax.

However, the Intermediate Form is not natural language, but a series of morpheme (and pseudo-morpheme) labels. Grasping the semantic context of the ambiguity can require a brutal mental effort.

Disambiguation After Synthesis of the Target Language

The most natural place to do disambiguation is on the synthesized target text. A tool named **WordPick** has been developed for this very purpose. (WordPick is a set of Microsoft® Word macros and styles, and accompanies the *CarlaStudio* software.)

Just the intuition of a target-language speaker can in many cases eliminate ambiguities. For example, suppose the following English sentence had been generated, with a gender ambiguity between *himself* and *herself*. This ambiguity comes through marked in *CarlaStudio*'s percent format, as:

```
John cooked breakfast for %2%himself%herself%.
```

Here is an enlarged view of how the WordPick tool presents this:

John cooked breakfast for %2% himself % herself %.

(WordPick formats the percent markings in a small font size and a different color, such that they serve only as slight visual separators.)

When the target-language speaker clicks on the appropriate box, the other option and the percent markings are removed:

John cooked breakfast for himself.

In some cases, language intuitions alone are insufficient, as different choices may be linguistically valid, but semantically different. For example,

Mary hit John, and %3% he she they left.

Here, the person disambiguating the text must determine which semantic alternative would have been used by the author, had that distinction been required in the source language.

Thus it can be seen that manual disambiguation can be a fairly involved process, most efficiently done once, rather than repeated for each target language.

8.2. Disambiguation Ideals

The following can be recognized as ideals for disambiguation:

- The effort of disambiguation should not be duplicated on multiple target languages.
- Disambiguation should be performed on natural language, not on a sequence of morpheme labels, such as on the Intermediate Form.
- It is preferable to have disambiguation performed by a native speaker of a target language for whom the distinctions are reflected in the language, rather than by the speaker of the source language, for whom the distinctions are not reflected in the language.
- Any disambiguation that can be performed by automatic processes should be done by such processes rather than manually.

8.3. The Disambiguation Cycle

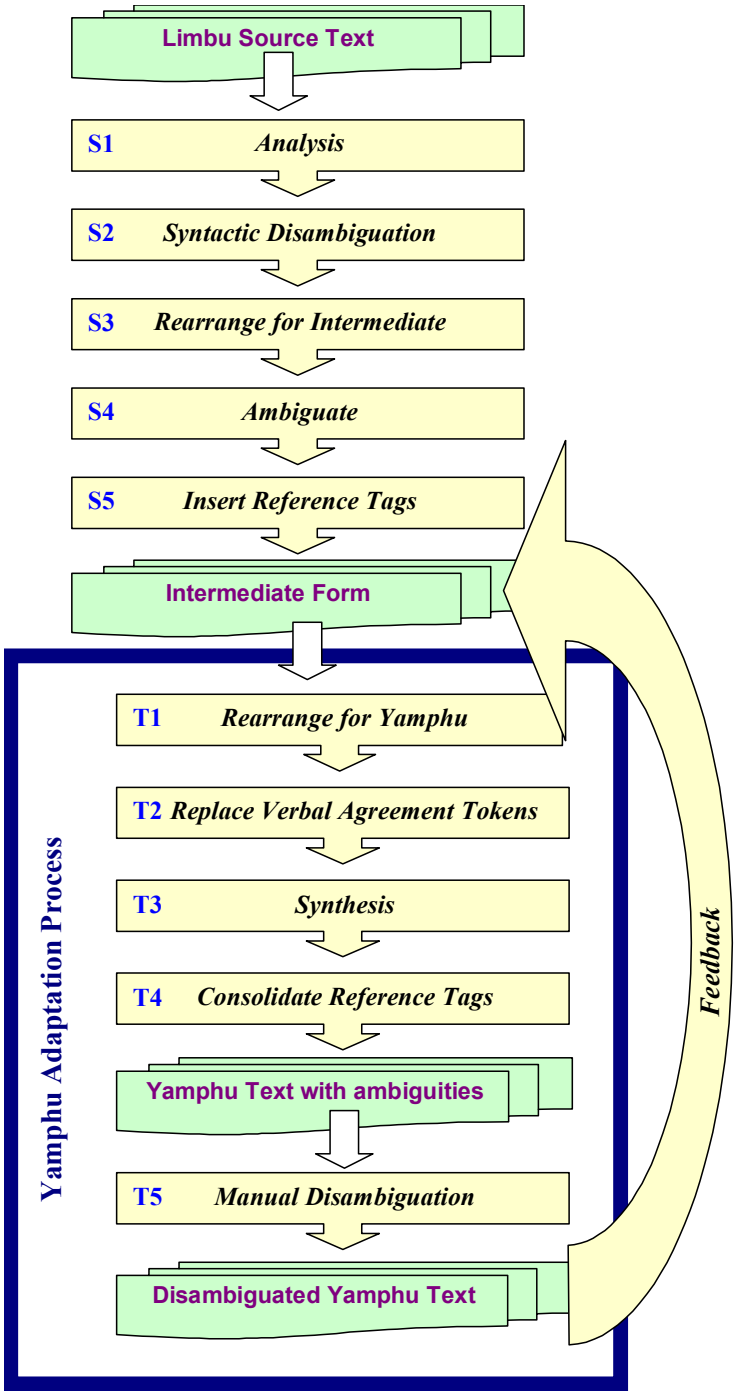
This study has developed a system of handling ambiguities in a manner that accommodates the above ideals. This system is thus an innovation.

The essence of this system is that each ambiguity in the Intermediate Form is tagged in such a way that the tag remains with the text, through all the successive processes. The selection made in WordPick on the synthesized target text serves as *feedback* to the Intermediate Form, which is thereby made unambiguous. Thus, when the text is adapted into other target languages, it will not be necessary to duplicate this same disambiguation.

8.4. Incorporating Disambiguation into the Adaptation Process

To understand how the disambiguation system works, an overview of the whole adaptation process is necessary. Omitting some small processes for clarity, the adaptation process can be represented as in Figure 14 below.

Figure 14 Flow of Adaptation Processes



❖ Source Process 1: *Analysis*

The analysis process parses the Limbu source text into morphemes.

Morpheme co-occurrence rules can be configured to select or reject analyses at analysis time. For example, /ha?/ is either the plural marker PL (when

suffixed to a noun) or the noun glossed **tooth**, in other contexts. A morpheme co-occurrence rule may specify:

tooth / [*n*] ~ __ i.e. **tooth** *is never suffixed to a noun.*

That is, an analysis of the compound noun **fairy-tooth** will be rejected, leaving only the analysis of **fairy** -PL. On the other hand, if the morphemes were in the other order, the compound **tooth-fairy** would not be rejected.

If the analysis is still ambiguous, each alternative is represented using *CarlaStudio*'s percent format.

For example, the Limbu verb root /cok/ is ambiguous, and could either be the verb glossed **do** or the verb glossed **be.desc**. Thus the word /cokki/ will be passed to the next process as these ambiguous alternatives:

%2% <*vi* **be.desc**> SIM % <*vt* **do**> SIM %

The /-ki/ suffix (labeled SIM) may legitimately co-occur with either verb, so morpheme co-occurrence rules cannot resolve this ambiguity.

❖ Source Process 2: Syntactic Disambiguation

In many cases, ambiguities in the analysis may be resolved syntactically. Syntax-based rules may either accept or eliminate certain analyses. To take the example of the verb /cok/ a step further, we may create a rule that says:

When /cok/ *follows a noun, accept the analysis of* **do**.

So for the text fragment /ya:mbɔk cokki/ (where /ya:mbɔk/ has been recognized as <*n* **work**>), the <*vt* **do**> SIM analysis will be accepted for /cokki/.

❖ Source Process 3: Rearrange for Intermediate Form

Unambiguous Rearrangement

The next process allows for an adjustment of the analysis, rearranging and replacing morphemes as necessary to shape the text into the structure of the Intermediate Form.

This is where the idioms described in Section 7.3.2 are converted to their functional equivalents, including compound nouns and idiomatic verbs.

More significantly, this is also where the functemization described in Section 7.3 occurs. The combination of a particular "chunk" of agreement suffixes with a particular "chunk" of agreement prefixes, perhaps also in combination with the free-standing agreement morpheme 2▶1 –all taken together– identifies a particular agent acting on a particular patient in a particular tense.

For example, a rearrange rule may say:

Wherever this morpheme pattern is found: 2▶1 2A- <*vt*> -DPS

replace it with: <*vt*> -P -PT -2A -dA -1P –sP

Thus, /yapmi kɛ-hip-si/, '*you(dual) hit me*' which is analyzed as:

yapmi	kɛ-	hip	-si
<*n* 2▶1>	2A	<*vt* **hit** >	DPS

is rearranged to tokenize the linguistic functions as:

<*vt* **hit**> P PT 2A dA 1P sP

There is no ambiguity in this, as /yapmi kɛhipsi/ always refers to 2d→1s actants.

Ambiguous Rearrangement

However, about half of the Limbu agreement forms have some kind of ambiguity, most often in a lack of distinction between dual and plural number or between past and non-past tense. Such ambiguity needs to be

carried through to the intermediate form, as certain target languages will indeed make such distinctions.

The normal means of introducing ambiguities into the analysis is to have two or more morphemes in the Limbu lexicon that both have the same morpheme form.[11] For example, recall the example of **do** and **be.desc**, which are both of the form /cok/. We might attempt to similarly duplicate one morpheme into two homophonous morphemes with different glosses.

For example, the /mɛ-/ prefix is currently analyzed as one morpheme glossed 3NSA (3rd person non-singular agent), as in /mɛhiptaŋ/, *they(dual/plural) hit me.*

Actants:		3d/p → 1s	
Limbu morphemes:	mɛ-	hipt	-aŋ
Morph glosses:	3NSA-	<*vt* **hit**>	-1

However, we might posit two separate morphemes, both with form /mɛ-/, one of which was glossed 3DA (3rd person dual agent) and the other of which was glossed 3PA (3rd person plural agent). The analysis would thus be ambiguous:

Analysis:	%2%	3DA	<*vt*	**hit**>	1					%	3PA	<*vt*	**hit**>	1						%					
Functemes:	%2%	<*vt*	**hit**>	P	PT	3A	**d**A	1P	sP	%	<*vt*	**hit**>	P	PT	3A	**p**A	1P	sP	%						

Indeed, this is the desired outcome.

The problem, however, is that the /mɛ-/ morpheme is not *always* ambiguous. For example, /mɛ- hip -sigɛʔ/ unambiguously denotes 3**p**→1d action, and cannot mean 3**d**→1d. The problem is even more accentuated in cases where the referential pattern of the agreement morphemes is more complex and irregular, as discussed in Section 5.3. This approach for introducing

[11] For example, this is the approach used by Watters where the genitive and ergative markers were homophonous in the source language, but not so in the target language.

ambiguities will not work for the function tokens, because the identification of morphemes is truly unambiguous.

Ambiguity Flags

The solution this study has developed relies on the Rearrangement process to specify one particular meaning, and to additionally insert *ambiguity flags*, pseudo-morphemes that tag a verb to indicate something of the alternative meanings.

To continue the above example of /mɛhiptaŋ/, the rearrangement rule would be:

Find: 3NSA <*vt*> 1

Replace with: <*vt*> P PT 3A pA 1P sP !da

The selection of the pA token indicates that the agent is **plural**, but the addition of the !da ambiguity flag indicates that the agent could have alternatively been **dual**.

The rearrangement rule might provide more than one ambiguity flag. For example, /hiptuŋsiŋ/ is ambiguous both for tense and for patient number. Any of these four meanings is possible:

1s→3**d** Past	1s→3**d** Non-Past
1s→3**p** Past	1s→3**p** Non-Past

Thus the replacement introduces two ambiguity flags into the analysis:

Replace with: *vt* -P -PT -1A -sA -3P -pP -!NP -!dp

The PT token indicates that the tense is **Past Tense**, while the !NP ambiguity flag indicates that the tense could have alternatively been **Non-Past**.

Likewise, the pP token indicates that the patient number is **plural**, while the !dp ambiguity flag indicates that patient number could have alternatively been **dual**.

This kind of Cartesian logic will account for most ambiguity patterns, but not all. For example, /yapmi kɛhip/ can refer to five logical possibilities:

	2s→1d	2s→1p
2p→1s	2p→1d	2p→1p

For such a pattern, a special ambiguity flag that refers solely to this particular pattern must be used. Thus the four tokens that specify the person and number of the agent and patient are somewhat redundant, but specified anyway so that the "syntax" of our protocol is not violated:

Replace with: *vt* -P -PT -2A -pA -1P -pP -!5p

This approach of introducing ambiguity flags into the analysis (as well as the approach of introducing function tokens unrelated to any real morpheme structure) is an innovation of this project.

❖ Source Process 4: Ambiguate

The next process converts the ambiguity flags into *CarlaStudio*-style ambiguities. This process is implemented as a program in the Perl scripting language. (Perl is arguably the world's most powerful language for text manipulation, hence its selection throughout this project.)

Input: *vt* P PT 3A pA 1P sP !da

Output: %2% <*vt* **hit**> P PT 3A dA 1P sP % <*vt* **hit**> P PT 3A pA 1P sP %

Each of the defined ambiguity flags expands out the provided tokens, or in the case of the special ambiguity flags like !5p, generates the appropriate set of alternatives.

It is possible that the word passed to this process is already ambiguous, such as with homophonous verbs. In this case the ambiguity flag is found within one or more of the provided alternatives, and is expanded out accordingly. For example, the sequence /lɛkkʰ/ represents two homophonous roots, /lɛkkʰ/ *'slap'* and /lɛkkʰ/ *'mislead'*. Used in the construction /yapmi kɛlɛkkʰaŋ/, this two-way verb ambiguity of verb meaning is in conjunction with the five-way referential ambiguity flag:

(28) Analysis of / yapmi kɛlɛkkʰaŋ /

%2% <*vt* **slap**> P PT 2A pA 1P pP !5p % <*vt* **mislead**> P PT 2A pA 1P pP !5p %

The Ambiguate process multiplies this out into a ten-way ambiguity:

(29) Result of the Ambigute process on the analysis of (28)

%10%	<*vt* **slap** >	P	PT	2A	**s**A	1P	**d**P	%
	<*vt* **slap** >	P	PT	2A	**s**A	1P	**p**P	%
	<*vt* **slap** >	P	PT	2A	**p**A	1P	**s**P	%
	<*vt* **slap** >	P	PT	2A	**p**A	1P	**d**P	%
	<*vt* **slap** >	P	PT	2A	**p**A	1P	**p**P	%
	<*vt* **mislead** >	P	PT	2A	**s**A	1P	**d**P	%
	<*vt* **mislead** >	P	PT	2A	**s**A	1P	**p**P	%
	<*vt* **mislead** >	P	PT	2A	**p**A	1P	**s**P	%
	<*vt* **mislead** >	P	PT	2A	**p**A	1P	**d**P	%
	<*vt* **mislead** >	P	PT	2A	**p**A	1P	**p**P	%

❖ Source Process 5: Insert Reference Tags

The final process performed on the text before it is considered completed Intermediate Form is to insert *reference tags*. A Perl script again accomplishes this. The purpose of reference tags is to uniquely identify each ambiguous word in the text, and each alternative within that set. Each set is assigned a number, and each alternative within that set is assigned a letter. This number and letter combination is tagged to the end of each alternative with a colon. For example:

(30) Insertion of reference tags

%2% <*vi* **be.desc**> SIM **:1a** % <*vt* **do**> SIM **:1b** %

%5% <*vt* **hit**> P PT 2A sA 1P dP **:2a** %
<*vt* **hit**> P PT 2A sA 1P pP **:2b** %
<*vt* **hit**> P PT 2A pA 1P sP **:2c** %
<*vt* **hit**> P PT 2A pA 1P dP **:2d** %
<*vt* **hit**> P PT 2A pA 1P pP **:2e** %

These tags are not a built-in part of *CarlaStudio* syntax, but our own extension of *CarlaStudio*'s syntax of ambiguity marking. (*CarlaStudio* treats them simply as morphemes.)

After this point, the text is considered to be in the final Intermediate Form, and ready to serve as a starting point for adaptation to any Kiranti target.

❖ Target Process 1: Rearrange for Yamphu

The first process on the target-side of adaptation is rearranging the Intermediate Form into a Yamphu-specific arrangement. Structural adjustments can be made, as well as adjustments to the glosses. For example, in Yamphu, the verb negation morpheme is a prefix in the Past Tense. This rearrangement can be effected with a rule like this:

v -N → NEG- *v* / __ -PT

❖ Target Process 2: Replacement of Verbal Agreement Tokens

This process is the reverse of the tokenization of the agreement morphemes. This process looks up in a table[12] the Yamphu morpheme structure for a given agent/patient combination. For example, a fragment of this table might look like this:

1A sA 3P sP → ▶3 1▶3
1A sA 3P dP → ▶3 1▶3 3DP 1▶3

[12] We use a *CC Table* with the *Consistent Changes* program to accomplish this.

1A sA 3P pP → ▶3 1▶3 3DP 1▶3

Multiple combinations may collapse into identical forms here.

❖ Target Process 3: Synthesis

The morpheme names, such as 3DP, are replaced with the surface form. The synthesis process may be required to select the appropriate allomorph for the phonological context, such as /ji/ instead of /ci/ because of voicing assimilation.

The result is a text file (not a *CarlaStudio* analysis file) in Yamphu. Ambiguities remain marked with the percent notation.

Much could be said about this process, but as far as our disambiguation cycle is concerned, the important thing to note is that "morpheme names" not found in the Yamphu lexicon are passed directly through. Our reference tags are conveniently considered to be such morphemes. For example:

Input: %2% <*vt* **see**> Du :3a % <*vt* **see**> 3P :3b %

Output: %2% kʰaksaji:3a % kʰaksami:3b %

On a technical side-note, because the reference tags are considered by *CarlaStudio* to be some word-final morpheme, their presence may inhibit the surfacing of allomorphs that are defined to exist in word-final environments. A small workaround is necessary to compensate for this. If the environment for an allomorph specifies *'where preceding a word boundary,'* an additional environment for that allomorph should be generated as *'where preceding a colon.'*

Original Environment: / _ # *before a word boundary*

Additional Environment: / _ : *or before a colon*

❖ Target Process 4: Consolidate Reference Tags

It is possible that ambiguities that arose because the Limbu verbs were homophonous may be collapsed by homophony between the Yamphu verbs. This is especially likely where there is a semantic link between the homophonous verbs. For example, the Limbu verbs that are glossed in English as *pour* and *upend* are homophonous. Indeed, one might even consider them to be different senses of the same Limbu verb, as there is a clear semantic link. However, for general Kiranti transfer purposes we treat them as separate verbs, as not all Kiranti languages can be expected to use the same verb for both purposes. If Yamphu also uses a single form that is semantically analogous to the Limbu verb glossed *pour/upend*, the synthesis stage will generate surface forms that are identical, except for their reference tags.

Similarly, as has been noted, during the process to replace verbal agreement tokens, collapsing of agreement ambiguities is frequent. When this happens, the Synthesis process will accordingly generate alternatives that are identical, except for their reference tags.

For example, suppose the Intermediate Form contains this five-way agreement ambiguity:

%5% <*vt* **see**> P PT 2A sA 1P dP **:47a** %
<*vt* **see**> P PT 2A sA 1P pP **:47b** %
<*vt* **see**> P PT 2A pA 1P sP **:47c** %
<*vt* **see**> P PT 2A pA 1P dP **:47d** %
<*vt* **see**> P PT 2A pA 1P pP **:47e** %

By the time the Synthesis process is finished with it, it looks like this:

%5% kʰaksa**:47a** %
kʰaksa**:47b** %
kʰaksaniŋ**:47c** %
kʰaksaniŋ**:47d** %
kʰaksaniŋ**:47e** %

Now the Consolidate Reference Tags process examines the alternatives in each ambiguity, and any that are identical except for the reference tags are consolidated by concatenating their identifying letters:

%2% kʰaksa**:47ab** % kʰaksaniŋ**:47cde** %

Of course, if the word is equally ambiguous in Yamphu as it was in Limbu, the ambiguity marking is completely removed, including the reference tags. For example, in Limbu, /hipsusigya/ is ambiguous as to whether the patient is dual or plural, so the Intermediate Form may look like this:

%2% <*vt* **see**> P PT 1A dA 3P **dP** **:32a** % <*vt* **see**> P PT 1A dA 3P **pP** **:32b** %

Once Yamphu synthesis is completed, we have no differences between these alternatives:

%2% kʰaŋʔinjuŋjiŋ**:32a** % kʰaŋʔinjuŋjiŋ**:32b** %

So reference tag consolidation makes that simply:

kʰaŋʔinjuŋjiŋ

without any ambiguity marking or reference tags. Yamphu has no disambiguation to offer the Intermediate Form on such words.

The text is now a readable Yamphu text that contains ambiguities.

❖ Target Process 5: Manual Disambiguation

Manual disambiguation can now be performed on the text, using a version of the WordPick tool that has been enhanced to recognize and appropriately

handle our reference tags. (Incidentally, WordPick works within Microsoft® Word, so the Yamphu-speaker performing the manual disambiguation can also make any other necessary adjustments to the text at the same time.)

As discussed in Section 8.1, WordPick displays the alternatives as "buttons". Our reference tags are formatted like the ambiguity markers rather than like the ambiguous forms themselves. For example, suppose WordPick is provided with this underlying plain text:

```
%2%kʰaksa:47ab%kʰaksaniŋ:47cde%
```

Here is an enlarged view of this text after WordPick has formatted it:

If the user clicks on the second button, the text is left as:

Thus the reference tag of the selected alternative remains in the disambiguated Yamphu text.

❖ Feedback Process

The final process takes these reference tags from the disambiguated text, and uses them to remove alternatives from the ambiguities in the Intermediate Form. (Once again, a Perl script performs this manipulation.)

Continuing the above example, since the reference tag **:47cde** is found in the disambiguated text, ambiguity #47 in the Intermediate Form is reduced to:

%3% <*vt* **see** > -P -PT -2A -pA -1P -sP **:47c** %
<*vt* **see** > -P -PT -2A -pA -1P -dP **:47d** %
<*vt* **see** > -P -PT -2A -pA -1P -pP **:47e** %

In cases where the reference tag has only one letter code (that is, where the selected alternative was not a consolidation of multiple alternatives) the

ambiguity in the Intermediate Form text is completely removed. For example, if the Intermediate Form text contained this ambiguity:

%2% < *vi* **be.desc** > SIM **:1a** % < *vt* **do** > SIM **:1b** %

and the disambiguated target text contained:

gardai:1b

then this word in the Intermediate Form will be reduced to the unambiguous:

< *vt* **do** > SIM

Of course, this Feedback process also removes the reference tags from the disambiguated target text to clean it up.

❖ Adaptation to Other Target Languages

Once the feedback process has removed ambiguities from the Intermediate Form, the text is better defined for adaptation into the next target language. Each successive target adaptation project will encounter fewer and fewer ambiguities requiring disambiguation.

9. OTHER ISSUES FOR ADAPTATION

We have thus established an architecture that embodies a strategy for transferring the results of disambiguation in one target language into all the other target languages. Furthermore, our architecture implements a hybrid interlingual strategy, based on the fact that Limbu "cares about" the same kinds of information being marked on the simplex verb as do the other Kiranti languages. Such features should be useful in any intra-Kiranti adaptation system. The next step is to broaden the scope of examination, to assess whether a sufficient degree of typological unity is also present in other areas.

9.1. Nominal Issues

Nominal morphology is simpler than the verbal morphology, so the hypothesis was that if we could get the verb to transfer, with all its affix slots, nominal morphology would be relatively easy. A somewhat complicating factor, it turns out, is the frequency with which complex inflected verb forms are nominalized, such that the nominal morphology is the outer layer around verbal morphology. Indeed, that is why we have category mapping strategies for analysis (as discussed in Section 2.4).

9.1.1. Case Marking

Ergative/Instrumental

The ergative and instrumental cases are marked identically in Kiranti languages (Ebert 1994:81). Van Driem (1987) goes to great pains to distinguish between ergative uses—marking the agent of a transitive verb—and instrumental uses—marking an instrument, cause, or means—in Limbu:

(31) LIMBU Ergative

mənа	**-lle**	co:g	-uba
man	**-ERG**	do	-3PT

'Someone has done it.'

(32) LIMBU Instrumental

a-	mik	-**le**	mɛn-	ni	-ʔe:	wa:	-ʔɛ
my-	eye	-**INS**	npG-	see	-npG	be	-1sPS/NPT

'I haven't seen it myself (lit. with my eyes).'

Of Yamphu, however, Rutgers (1998:58) states that "there is no formal criterion to distinguish an ergative from an instrumental case… Semantic and formal investigation… lead to only one suffix" which marks the same types of things as the ergative and instrumental markers in Limbu:

(33) YAMPHU Ergative/Instrumental (here marking the agent of a transitive verb)

la:ma	maʔ	-ye	lu:s	-u	Mo:gamm	-**æʔ**	aseʔŋa
Lama	not_be	-FCT	say	-›3	Mogamma	-**ERG**	yesterday

'Mogamma said yesterday the the Lama wasn't there.'

(34) YAMPHU Ergative/Instrumental (here marking a cause)

mo	-ba	kho	-eʔe	ã:r̂	-**dæʔ**	hago	khad	-iŋ	-ma	sum	-baŋ
that	-ELA	s/he	-POS	spirit	-**INS**	now	go	-EXPS	-12NS	three	-UN

'Thanks to his courage, we went on, the three of us.'

Here we have a clear case of Kiranti languages "caring about the same kind of stuff" in that they all mark ergative and instrumental cases. Furthermore, it is also a case of them having the same kind of stuff that they do *not* care about, in that none of them care to make a formal distinction between ergative and instrumental cases. These factors give a real boost to the feasibility of adaptation.

Absolutive

Rutgers (1998:57) says that the absolutive case is a suffix *with zero marking* in Yamphu, applied to patients of transitive verbs, and typically the subjects of intransitive verbs. Van Driem (1987:84) also describes an absolutive case in Limbu, which he says is overtly marked only when the noun is definite. In stark contrast, Ebert (1994:82) says that there is no absolutive marker in Kiranti languages, and that the Limbu marker is nothing more than a marker of definiteness. Thus, Kiranti adaptation does not need to deal with the absolutive case. The absolutive case is thus another feature that Kiranti languages are fairly united in *not* caring about, and so it makes no obstacle

for adaptation. (Definiteness, on the other hand, seems to be slippery in Kiranti, typically involving demonstratives. Limbu's suffixal definite marker is the exception.)

Genitive/Possessive

The genitive case indicates a belonging together or possession (in which case it marks the possessor, not the possessed). In Yamphu, the genitive <*-mi*> is not homophonous with the ergative case marker <*-lle*>.

(35) YAMPHU Genitive

namba	-ji	**-mi**	jimma
father-in-law	-NS	**-GEN**	land

'father-in-law's land'

In Limbu, however, the genitive case marker is almost identical to the ergative/instrumental case marker. (In just a few limited phonological environments, it surfaces slightly differently, and can be thereby be distinguished from the ergative marker.) This would initially seem to present a significant obstacle to adaptation from Limbu to Yamphu, where the two cases are very distinct. However, the Limbu language actually does have a strategy to prevent ambiguity between these homophonous cases, namely, the placement of a possessive prefix on the possessed noun:

(36) LIMBU Genitive

tumma	**-re**	**ku-**	sa
first_wife	**-GEN**	**her-**	child

'first wife's child'

In the Limbu analysis, a disambiguation rule states that if the ambiguous GEN/ERG marker is followed by a word containing a possessive prefix, resolve the ambiguity by accepting the GEN interpretation. (Transfer of the possessive prefix itself is discussed in Section 9.1.2.) Thus, much of the ambiguity arising from this ERG/GEN homophony can be easily resolved.

Somewhat more thorny, however, are the non-possession uses of the genitive described by van Driem, in which case the possessive prefix is dropped.

(37) LIMBU Genitive (non-possessive)

si: -**re** khɔrɛːŋ
wheat -**GEN** bread
'wheat bread'

To automatically resolve these type of uses would require a much deeper analysis. However, Webster reports that this construction 'sounded odd' to Limbu speakers he checked with, who wouldn't use a genitive here at all. What turns out to be a greater challenge is that, according to Rutgers, Yamphu makes a distinction between a "genitive" case <*-mi*> and a "possessive" case <*-æʔæ*>. I have not managed to pin down a semantic distinction between these; Clearly it is not related to the conventional distinction of possessive being a sub-category of genitive, as the "possessive" marker can be used in non-possessive contexts:

(38) YAMPHU

iskul -**i** -ha -ji -so jammai
school -**POS** -PLNR -NS -too all
'all the school children too' (lit. 'everyone of the school too')

Likewise, the "genitive" case can be used in semantic contexts that are indeed possessive, as in (35) above (i.e. "father-in-law's land").

It seems that both cases can function adnominally, that is, they can modify an overt head on which they are dependent; the genitive as demonstrated in (35), and the possessive as:

(39) YAMPHU Possessive

sya:l -**æʔæ** jal -di
jackal -**Pos** trick -EXH
'[It was] the jackal's trick' *(Rutgers 1998:447)*

Both cases can also occur as independent nominal heads:

(40) YAMPHU Genitive (Independent)

ikko thaʔma -**mi** nu:rok ceʔmælo ikko -**mi** maŋæ ceʔmællo
one old_woman -**GEN** well ploughing.REP one -**GEN** bad ploughing.REP
'I was allegedly ploughing one wife's [field] well and the other wife's [field] poorly.'

(41) YAMPHU Possessive (Independent)

k -**æʔæ**
I -**POS**
'mine'

One difference here, however, is that when used pronominally, POS attaches to the base form of the pronoun (as in (41)), while GEN attaches to possessive prefix:

(42) YAMPHU Genitive (Independent)

kaŋ.mi tu.ye
my.GEN be.FCT
'I have mine'

When what the independent genitive refers to is of dual number, the non-singular suffix <*-ji*> (NS) is added.

(43) YAMPHU Genitive (Independent, Dual)

kaŋ.min.ji
my.GEN.NS
'mine'

However, if the referent is of plural number, the possessive case must be used instead. In addition, the plural nominalizer <*-ha*> (PLNR) will occur, as the possessive requires this whenever the referent is of plural number.

(44) YAMPHU Genitive Replaced with Possessive

yaʔmi -**di** -**ha** na:nisa -ji
person -**POS** -**PLNR** sister -NS
'people's sisters'

Both the POS and GEN markers may be used in place of the Nepali genitive <*-ko*> in borrowed constructions, with no clear semantic or syntactic basis for the choice between them:

(45) YAMPHU Possessive (in borrowed Nepali genitive structure)

Nardajiemb -**æʔæ** **cheu** -**beʔ**
person -**POS** **side** -LOC
'to the side of Nardajiemba' *(Rutgers 1998:371)*

(cf. Nep. <X-*ko cheu-ma*> X-GEN side-LOC)

(46) YAMPHU Genitive (in borrowed Nepali genitive structure)

igh -a mo -dok -**mi** **la:gi** aŋ.ʔitt.u.ŋ.ha
this -PLNR that-like -**GEN** **sake** make.PF. →3.EXAG.PLNR
' I have made this [beer] for that reason ' *(Rutgers 1998:541)*

(cf. Nep. <X-ko lagi> X-gen sake 'For X's sake/purpose')

Thus, Yamphu appears to have two structurally distinct strategies for marking a genitive, and these are—to at least some extent—in competition. It may be that eventually, one strategy will completely supplant the other.

However, consider that English itself has two distinct structural patterns by which the genitive can be encoded (e.g. *the airplane's speed* vs. *the speed of the airplane*), and that some phrases are "more grammatical"[13] in one structure than the other (e.g. *the book of judgment* vs. **the judgment's book*; *the state of Montana* vs. **Montana's state*) or even that a switch of the genitive pattern may have semantic implications. (For example, *"Joe's accident"* is typically preferred to *"the accident of Joe"* as the latter may imply that Joe's very existence was a mistake.[14]) These two strategies are in some cases in competition, quite freely interchangeable (e.g. *'the room's furnishings'* vs. *'the furnishings of the room'*), and yet both strategies have a pretty stable hold in the English language.

One wonders if perhaps Yamphu's two structurally distinct strategies for the genitive operate in an analogous manner. Obviously, this is not because of any inheritance or borrowing between English and Yamphu. Rather, this is a reflection on the very character of the genitive, which can be incredibly multi-

---

[13] In every natural language, areas exist in which there is no clean dichotomy between grammatical and ungrammatical, but rather the issue is one of *degree* of grammaticality or naturalness: '*highly marked', 'awkward', 'marginal', 'only permitted poetically',* etc. This fact poses a challenge for transfer-based adaptation systems, which need to make a binary choice between either eliminating or retaining a hypothesized analysis. A strength of statistical machine translation is that an assessment of 'probable grammaticality' is treated as a non-discrete variable to be factored into an ultimate ranking, not merely a judgment of *valid* vs. *invalid*.

[14] Actually, in this kind of case, English allows us to avoid that ambiguity by applying the other genitive strategy on top of this one. e.g. *"that accident of Joe**'s**"* (but probably not *"*?the surface of the pool**'s**"*). In Yamphu, however, the simultaneous application of both genitive strategies appears to be unattested in Rutger's corpus.

functional, able to indicate not just possession, but also purpose, constitution, classification, measure, or perhaps virtually *any* relationship (Cunningham 2006).

In any case, this kind of alternation between Yamphu's POS and GEN markers is highly complex, and thus poses a serious challenge for automated transfer. Probably the best we can do is default to the "genitive" case, and convert that to the "possessive" case in contexts in which syntax demands it, leaving other contexts to be manually corrected by a Yamphu speaker.

Locative

Limbu's locative marker <*-ʔoː*> and Yamphu's locative marker <*-peʔ*> both mark both place and direction.

(47) LIMBU
nyaʔ -re -**ʔoː**
aunt -GEN -**LOC**
'to/at Auntie's place'

(48) YAMPHU
Guruŋ -dæʔæm -**beʔ**
Gurung -POS -**LOC**
'to/at the Gurung's place'

It is worth noting here that the Yamphu locative can attach only to the "possessive" <*-œʔœ*> (POS), not the "genitive" <*-mi*> (GEN). This, then, is one syntactic cue by which the appropriate POS/GEN choice can be inferred during the transfer process.

Comitative/Sociative

The Limbu case <*-nu*> that van Driem calls "comitative" is a clear cognate of the Yamphu case <*-nu ~ -nuŋ*> that Rutgers calls "sociative". Like the English gloss *"with"*, this case is used to indicate both accompaniment (*"meet **with** him"*) and instrumentality (*"cut **with** a knife"*). One difference is that in Limbu, when two nominals are thus joined, according to van Driem, verb agreement will be with the combined unit. Apparently in Yamphu, verb

agreement may optionally be with either the combined unit or with just one of the entities. (Presumably the subtle semantic difference is akin to the difference between *"I walked with my friend"* and *"My friend and I walked."*) Since the target language is the more flexible here, this is not a problem for adaptation.

Another difference, however, is that if the combined unit is to receive ergative marking, Yamphu will mark the combined unit, while, according to van Driem (1987:50), Limbu marks both constituents separately, as indicated with square brackets in the following examples:

(49) YAMPHU

[Nardajiemba	-**nuŋ**	Reliy]-æʔ	sa:ro	tha:ppuwa.be	seʔ.end.u.ji.ro	
[Nardajiemba	-soc	Rele]-ERG	very	fishline.LOC	kill.np.›3.3NS.REP	
	khem.mitt.w.e		læ:			
	hear.PF.›3.FCT NW					

'He had heard that Nardojiemba and Rele often went fishing with the fishing lines.'

(50) LIMBU

[syaʔl]-le	-nu	[ũʈh]-ille	soːʔl.in	yəllik	ceːsu	
[fox]-ERG	-COM	[camel]-ERG	sugar_cane.DEF	much	ate^{du}	

'The fox and the camel ate lots of sugar cane.'

The implication for adaptation is that in transfer to Yamphu, the sequence -ERG -COM can be reduced to -COM. Webster (p.c. 2006) reports, however, that Limbu speakers he checked with found the -ERG -COM construction odd, preferring the same structure as shown for Yamphu. In that case, the above reduction rule would go unutilized.

Mediative

Yamphu's mediative case *<-la ~ -lan>* patterns very much like Limbu's mediative case *<-lam>*, which, according to van Driem, derives from the same etymon as the Limbu noun *lam 'road'*. In both languages, the mediative is used to express the route or means by which an event occurs. When translated into English, the Kiranti mediative may result in wording like ***"by***

way of** [route X]"* or "from** [location X]*" or even "***in** [abstract medium X]*". However, usage seems to align fairly consistently between Limbu and Yamphu:

(51) LIMBU

peːnibaːn	-**lam**
Nepali	-**MED**

'in Nepali'

(52) YAMPHU

k.æk.ko	khasi.ha	khap	-**la**	-re	luː.jæ.n.u.ŋ.æ	læː
I.ERG.TH	Nepali.PLNR	language	-**MED**	-CEF	say.bring.NP.›3.EXAG.FCT	NW

'I see that I've been talking in Nepali'

The mediative case, then, bears out the axiom that the relatedness of the languages provides the most fundamental basis for adaptation.

Elative

Yamphu has an elative case *<-pa ~ -paŋ ~ -pan->*, marking the starting point in space (or by analogy, time) from which a departure occurs. Note that both types of departure are present in this example:

(53) YAMPHU Elative

mo	-**ba**	hoŋma	-**ba**	ka	ram -bug	-iŋ
that	-**ELA**	river	-**ELA**	I	walk -start	-EXPS

'Then (lit. 'from that [point in time]') I departed from the river'

Van Driem posits an elative case in Limbu, too, comprised of LOC + COM in alternation with LOC + MED.[15]

(54) LIMBU Elative

tɔŋba	nasi	thuŋ	-u	-ŋ	hɛkkɛlle	khɛŋ	cumluŋ	-ʔoː	-**lam**	pu	-eːkkeː
tungba	five	drink	-3P	-1sA	so	that	bazar	-**LOC**	-**MED**	bird	-like

pɛr	-aŋ	-ba
fly	-1sPS/PT	-IPF

'I drank five tungbas, so I flew back from that bazaar like a bird.'

Thus, in transfer to Yamphu, a rule can replace the sequence -LOC -MED or -LOC -COM with Yamphu's own case marker -SOC.

[15] Perhaps on this basis we may hypothesize that the Yamphu elative *<-pa ~ -paŋ ~ -pan->* is similarly derived from LOC *<-peʔ>* + MED *<-la ~ -lan>*.

Note also in (53) that *moba 'then' (lit. 'from that')* is a high-frequency word in Yamphu. It seems to correspond best with Limbu's *hɛkkyaŋ '[and] then'* (which is, in fact, the highest-frequency word in the Limbu corpus, as shown in Figure 8 on page 42), and is thus substituted at the word level. (Both are derived from distal roots. Indeed, Ebert (1994:93) lists several Kiranti languages for which the anaphoric discourse connectors glossed *'then, thereafter'* are derived from such roots.)

9.1.2. Possessive Prefixes

As shown in (36), Limbu marks a possessed object with a possessive prefix. In Yamphu, somewhat corresponding prefixes exist, but they are restricted to a handful of "person words" (*<namba> 'father-in-law', <laŋgam> 'friend',* etc.).

(55) YAMPHU

am- nisa
your- younger_sibling
'your younger sibling'

Ba:ja -mi **khom**- ba
Baja -GEN **his**- father
'Baja's father'

The possessive prefix in Limbu, on the other hand, does not have this restriction:

(56) LIMBU

siŋbo:ŋ -ille **ku**- bo:ŋ -ʔo:
tree -GEN **its**- base -LOC
'at the base of a tree'

It would seem that the class of words that can take the possessive prefixes varies somewhat among Kiranti languages. According to Sueyoshi Toba (cited n.d. in Ebert 1994), in Khaling, the possessive prefixes may attach to kinship terms (as in Yamphu) and also to terms for body parts. For adaptation purposes into a specific target language, this requires a transfer rule that deletes any possessive prefix that is not attached to a member of the target language's class of "possessable" nouns.

9.1.3. Numerals

Rutgers provides Yamphu numerals up to twenty, and then the decades up to ninety. However, above six, Nepali numerals have become the norm. This parallels the situation in Limbu. In both Yamphu and Limbu, the numeral 'one' may be used (as with Nepali's *euʈa* '*one*') as an indefinite article:

(57) YAMPHU

ikko	damai	-dæm	-beʔ	yu:s	-a -j	-iŋ
one	tailor	-POS	-LOC	descend	-PT -DU	-EXPS

'We descended to the house of a tailor.' *(Rutgers 1988:100)*

(58) LIMBU

anche:	anche:	mu	ya:kkha	-ʔo:	**lɔkthik**	sya?l	-**dhik**	mu	way -ɛ
before	before	REP	jungle	-LOC	**one**	jackal	-**one**	REP	be -PT

'Long ago there lived a jackal in the jungle.' *(van Driem 1987:345)*

The Limbu word *lɔkthik* (literally "*only one*") is a common emphatic form of the basic form *thik 'one'*.

9.1.4. Deixis and Vertical Location

The encoding of vertical space in Kiranti languages has often been remarked upon. This feature is fairly pervasive throughout these languages. In addition to the vertical dimension being indicated by adverbs, it may also be lexically incorporated into verbs, marked on demonstratives and pronouns, and marked directly on nouns as local case-marking.

Vertical Case-Marking

Vertical case-marking indicates that noun so-marked should be understood to be lying higher (UPW), lower (DWN), or on relatively the same horizontal plane (HRZ) as the speaker:

(59) YAMPHU

Rokhiemma	yoŋ	-æʔ	-**mu**	tu:-yag	-a
Rokhiemma	water	-POS	-**DWN**	be -stay	-PT

'Rokiema stayed at the place of the spring [which is lower than here].'

The vertical case markers applied to nouns are a fascinating area. However, Limbu is somewhat unusual within Kiranti in that the vertical case markers cannot be applied to nouns. (They can be used in all the other constructions where other Kiranti languages use them, such as with demonstratives and pronouns, where they have fixed lexical forms. They are not otherwise productive.) For the purposes of adaptation from Limbu, then, we do not need to address target-language vertical case-marking of nouns.

From Two Proximity Markers to Three

All Kiranti languages can mark the vertical dimension in the deictic system. Ebert (1994) claims that Kiranti languages have a two-category system from which demonstratives and adverbs are derived: proximal (*'this', 'here'*) and distal (*'that', 'there'*). (She notes that Toba (1984) had posited an additional far-distal term in Khaling, but dismisses it as a nominalized form of the 'same-vertical-level' distal term.) Rutgers, however, posits a three-category system in Yamphu, built around three demonstrative pronouns: *<igo> 'this'*, *<akko> 'that'*, and *<mo> 'that/yon'*. If Yamphu, then, differs from other Kiranti languages by having a three-way distinction of proximity, it introduces an issue for adaptation: Which Yamphu distal should be generated from the distal in the source language?[16]

---

[16] Opgenort (2005) presents the demonstrative system of Jero (a Western Kiranti language) as being based on a pattern of five bound deictic morphemes. Implicit in this presentation is the existence of three degrees of proximity, of which the furthest degree is obligatorily marked for vertical case:

<a-> 'near (near the speaker)'
<u-> 'distal (near the hearer)'
<nɔ-> 'yonder (at the same elevation)'
<tɔ-> 'yonder (up)'
<yɔ-> 'yonder (down)'

As the Jeru vertical case markers are <-na>, <-ta>, and <-ya> for 'same level', 'up', and 'down' respectively, if we were to suppose that Jero's distal actually marked 'distant from the speaker' (as opposed to 'near the hearer' as Opgenort has labeled it), an alternative explanation would suggest itself, namely that Jero actually distinguishes only two degrees of proximity, and that only the distal has the option of accepting marking of the vertical deictic, resulting in the five base forms on which Jero's demonstrative system patterns. Somewhat further afield in Kham, Watters (2002) posits three degrees of proximity: 'proximate', 'distal

Hart (2004), which is an analysis of the vertical encoding of Yamphu as described in Rutgers (1998), puts it this way: "Yamphu demonstratives distinguish three degrees of distance: proximal (here), distal (there), and far-distal (way over there)." In a table of comparative Kiranti vertical markings on the demonstratives, the form that Hart aligns with the distal form of other Kiranti languages is Yamphu's *<akko>* form (that he calls "distal"), not Yamphu's *<mo>* form (that he calls "far-distal"). For adaptation purposes, however, we must be very careful about alignment issues. First, it should be noted that Rutgers has not explicitly stated that *<mo>* is more distant than *<akko>*; the label "far-distal" is not his. I have not been able to identify any significant semantic distinction in the contexts in which each is used. The two distal markers may, in fact, be so semantically overlapping as to be in competition. Indeed, the two distal categories conflate in the demonstratives of relative place and direction. Second, as Hart himself notes, *<mo>* occurs in the texts far more frequently than *<akko>*. For that matter, recall that *<mo>* is the distal root from which the high-frequency discourse connector *moba* is derived (cf. p. 102). Finally, *<mo>* would seem to be a cognate to certain other Kiranti distal markers, while this does not seem to be the case for *<akko>*, as may be observed in Figure 15 below.

(within view)', and 'remote', but instead these deictics being of cross-indexed by vertical deictics, these three deictic primitives are suppemented by 'up', 'down', 'front', 'back', 'right', 'left', and 'where', which together form the set of ten deictic primitives on which various deictic expressions may be based. As Watters (2006) puts it, "It is in the use of 'vertical orientation' suffixes that Kham departs from the Kiranti languages. In Kham, vertical orientation is expressed only through deictic primitives."

Figure 15 Kiranti Demonstrative Roots

	Proximal	Distal	
Limbu	kɔt	$k^hɛt$	
Bantawa	o	mo	
Camling	o / u	tyo/tyu	
Thulung	o	mö	
Khaling	tä	mä	
Dumi	tom	mom	
Yamphu	igo	akko	mo

(Data from Ebert 1994:91 and Rutgers 1998:94)

Thus, for adaptation purposes, we transfer the Limbu distal forms to the Yamphu distal forms that derive from <*mo*>, and we never generate forms based on <*akko*>.

Specificity of Deictic Location

One other wrinkle for adaptation arises in that Yamphu has two variants of the basic demonstratives: <*igo*> *'this' and* <*akko*> *'that'* also have shorter forms <*i*> *'this' and* <*ak*> respectively in locative contexts. (<*mo*> has just one form.)

Figure 16 Yamphu Basic Locative Demonstratives

***dem.* + LOC**	**Location/Direction**	***dem.* + ELA**	**Origin**
igo.beʔ ~ i.beʔ	'to/at this place', 'here'	igo.ba ~ i.ba	'from here'
akko.beʔ ~ ak.peʔ	'to/at that place', 'there'	akko.ba ~ ak.pa	'from there'
mo.beʔ	'to/at that place', 'yonder'	mo.ba	'from there', 'then'

(Rutgers 1998:97, morphology added)

Although often used interchangeably with the shorter forms, the longer forms have a subtle semantic difference: They refer to a specific spot, while the shorter forms refer to a more general location. For this reason, it seems possible that, although Rutgers does not analyze this as a distinct morpheme, the <*-ko*> observed in both the longer forms derives from some specificity or emphatic marker. Indeed, Yamphu's theme marker is also <*-ko*>, so it may even be possible that these two morphemes derive from a common etymon. I would not take that so far as to synchronically call them the same marker.

Rutger's texts attest the word *igo.go* (this.TH), in which the theme marker has been placed on the absolutive form of the proximal demonstrative pronoun (e.g. Rutgers 1998:449). When used independently as absolutive or ergative demonstrative pronouns (i.e. not in the locative context where specificity may be relevant), the longer <-*ko*> form is required:

Figure 17 Yamphu Demonstrative Pronouns

	ABSOLUTIVE	PLURAL	ERGATIVE
THIS	i**go**	i**g**ha	i**go**sæʔ ~ i**gw**eʔ
THAT	ak**ko**	ak**k**ha	ak**ko**sæʔ ~ ak**ko**eʔ ~ ak**kw**eʔ
THAT/YON	mo	moha	mosæʔ ~ moeʔ ~ mweʔ

(Rutgers 1998:94, emphasis added)

To analyze both <-*ko*> morphemes as synchronically identical results in *igogo* being parsed as 'this.TH.TH', which seems to fail to capture the different roles currently played by the morphemes, the first of which seems more tightly bound to the deictic root.

In any case, the real issue for adaptation is that Limbu's locative deictics (*kəʔ-oː* {this-LOC} *'here';* and khɛʔ-oː {that-LOC} *'there'*) do not make this same subtle distinction between general and specific location. It appears that the more frequent locative form in Yamphu is *ibeʔ* (not *igobeʔ*), and since this is also the less-marked form, this is the form we shall select in the Yamphu transfer process for a Limbu analysis of {this -LOC}.

Vertically-Specified Demonstratives

In Kiranti languages, demonstratives can specify not only a referenced object's proximity to the speaker, but also its vertical level relative to the speaker.

Figure 18 Kiranti Vertically-Specified Demonstratives

	Limbu	Bantawa	Yamphu
'over here'	**kɔt**.*na*	**o**.*du*	**i**.beʔ.*yu*
'up here'	**kɔt**.*thoː*	**o**.*yu*	**i**.bet.*tu*

'down here'	**kɔt**.*yo:*	**o**.*ya*	**i**.be?.*mu*
'over there'	**khɛt**.*na*	**mo**.*du*	**mo**.be?.*yu*
'up there'	**khɛt**.*tho:*	**mo**.*yu*	**mo**.bet.*tu*
'down there'	**khɛt**.*yo:*	**mo**.*ya*	**mo**.be?.*mu*

(Data from Rutgers 1998:97, morphology added, Ebert 1994:91, and Webster p.c. 2006)

In every other Kiranti language on which I have data, these locative demonstratives are constructed of {proximity deictic} + {vertical deictic}. In Yamphu, however, the locative suffix <-*pe?*> intervenes (not that this is any obstacle for adaptation).

Once again we do find in Yamphu some distinctions not present in Limbu. Again it seems that the contrast between the short and long deictic stems—that is, the absence or presence of the mysterious <-*ko*> morpheme—may play some role in this:

Figure 19 Yamphu Demonstratives of Place and Direction

Place *dem* + LOC + *level*		**Direction/Place** *dem* + /**ko**/ + /i?/ + *level*
i.be?.*yu*	'[over] here'	i.**g**.i?.*yu*
i.bet.*tu*	'up here'	i.**g**.in.*du* ~ i.**g**.it.*tu*
i.be?.*mu*	'down here'	i.**g**.im.*mu* ~ i.**g**.i?.*mu*
ak.pe?.*yu*	'[over] there'	ak.**k**.i?.*yu*
ak.pet.*tu*	'up there'	ak.**k**.it.*tu* ~ ak.**k**.in.*du*
ak.pe?.*mu*	'down there'	ak.**k**.i?.*mu* ~ ak.**k**.im.*mu*
mo.be?.*yu*	'[over] there', 'yonder'	m.i?.*yu* ~ m.i.*yu*
mo.bet.*tu*	'up there/yonder'	m.it.*tu* ~ m.in.*du*
mo.be?.*mu*	'down there/yonder'	m.i?.*mu* ~ m.im.*mu*

(Rutgers 1998:97, morphology added)

Here, the series on the left, which indicates a place of particular proximity and relative vertical plane, is derived from the shorter demonstrative roots. The series on the right, which may also indicate place but is primarily used for indicating direction, is derived from the longer demonstrative stems that contain <-*ko*>. (Some other older morpheme <-*i?*> is also evident in this

derivation, but I have no theory yet as to its origin.) Again we will default to the less-marked forms of the left column, as indicated in Figure 18.

9.2. Verbal Issues

9.2.1. Verbal Complements

According to Doornenbal (2004), all Limbu verbs roots are monosyllabic. That is, the verbal prefixes (such as for agreement and negation) and verbal suffixes (such as for agreement, tense, and reflexivity/reciprocity) attach to a single root syllable. Since there are a limited number of phonotactically-valid syllables, how can these represent an open class of verbs? Limbu accomplishes this by allowing what Weidert (1985) calls "optional extensions in pre-verbal head position." That is, the verb may prefix an additional argument to complete its semantic package. For example, combining a head of *<sen> 'inquiry'* with the verb root *<dos ~ do> 'do'* results in the verb *sendoma 'to ask'*.

(60) sen- dos -u
inquiry do -3sPT
'he asked him'

(61) sen- mɛ- dos -usi
inquiry 3ns- do -3nsPT
'they asked him'

Van Driem (1987:367) presents the verb from a somewhat more synchronic perspective when he says that prefixes follow the first syllables of a polysyllabic verb, to attach to the final "core" syllable. However, in the light of Givón's (1971) adage that "today's morphology is yesterday's syntax", it seems clearer to describe the polysyllabic verb in terms of an incorporated object.

Historically, this would seem to have developed from a system similar in some respects to that of Nepali, in which a smallish set of verbs (with general

meanings glossed *'to do', 'to fall', 'to attach'*) are productively paired with a wide variety of nouns and sometimes adjectives. For example,

(62) kurɑ gər -nu
item/matter do -INF
'to converse/discuss'

In the Limbu examples of (63) and (64), however, the direct object has been truly incorporated, and is perceived by speakers as an integral part of the verb. Furthermore, it is not available for the normal nominal morphology, such as for number or case marking. In verbs like this, the semantics are derived primarily from the incorporated object, and the verb root serves merely as a convenient parking place for inflection (Doornenbal, p.c. 2006). There is a also smaller set of other Limbu verbs in which it is the incorporated object that seems to bring little in the way of a new semantic contribution (Doornenbal 2004). For example:

(63) wa- hɔp -siŋ -ma
water wash -REFL -INF
'to wash oneself'

(64) iŋ- dɔŋ -ma
thing agree -INF
'to agree'

Object incorporation is apparently common across Kiranti languages, but the issue for adaptation is that different languages do not necessarily use object-incorporated verbs in the same way. For instance, for expressing a particular semantic notion, the combination of object and verb root used may be a rather idiomatic convention, unique to that language. Moreover, a semantic notion that is expressed using an object-incorporated verb in Limbu may be expressed by a simple verb in another Kiranti language. For example, Limbu's object-incorporated verb *sen.do.ma 'to ask'* (exemplified in (60) and (61)) is the semantic equivalent of Bantawa's simple-root verb *sen.ma 'to ask'* (in which the root is clearly a cognate of Limbu's incorporated object) and likewise of Yamphu's single-root verb *sim.ma 'to ask'*.

Again our function-over-form approach allows us to extend the reach of adaptation. The Intermediate Form, then, as a representation of an idealized Limbu analysis, should not reveal the object-incorporated structure of the original morphemes. Rather, it should contain a unique verb identifier (such as *ask* in the case of the current example) for each object-incorporated verb. Each Kiranti target language, then, can begin transfer from that consistent form, and use either a simple or object-incorporated structure as appropriate for that verb in that target language.

9.2.2. Nominalization

Nominalization is a pervasive feature of Kiranti languages, seemingly applicable to virtually every part of the language, even to whole sentences. The semantics of such nominalizations have been described by the authors of the various grammars in extremely divergent ways, in some cases, seemingly directly contradicting each other. For example, the nominalization of a stand-alone clause in Yamphu is described as marking 'background information', while Bickel (1999, cited in Watters 2006) describes the semantics of a similar structure in Belhare as having "an intrinsic potential for controversy", the opposite of background-marking. Watters (2006) builds the case that both are actually part of a larger, typologically-unified system. Typological unity—that is, languages "caring about the same kind of stuff"—is what we like to see for adaptation purposes, but there is no doubt that structurally, Kiranti languages go about nominalization by means that are often significantly divergent.

Some commonality of structure is present, too, of course. Virtually all Kiranti languages seem to have inherited the nominalizer *<-pa>* in some fashion:

(65) LIMBU

ku-lum-ʔoː	mɛ-bhaŋ-u-**ba**	way-ɛ
its-between-LOC	nsAS-fence.off-3p-**NOM**	be.PT

'In between there was a separating wall they had built.'

(66) YAMPHU (here nominalizing a borrowed Nepali verb *makinu 'to mold'*.)
mak-**pa** li:-ghad-a
to.mold-**NOM** become.PROC.PT
'It's completely molded away.'

However, this same etymon may be used in extremely different ways. In Yamphu, *<-pa>* suffixes only to loan verbs, and never to a finite form (as seen in (65) in Limbu), and furthermore, this nominalized loan may only be used as part of a periphrastic construction that involves the Yamphu verbs *<læ:t> 'to do'* or *<lis> 'to become'*, where the choice between these is determined by the loan verb's transitivity. Thus, its usage is highly restricted in Yamphu.

In Limbu, on the other hand, *<-pa>* is a general-purpose nominalizer that is utilized in all of Limbu's nominalization strategies. The only divergence from this pattern is that the construction of the active participle additionally requires the prefix *<kɛ->* as in *<kɛ-sep-pa> 'he who kills'* (Watters 2006).

The existence of an active participle is a shared feature of all Kiranti languages, although its construction varies. Yamphu utilizes a specialized marker for this purpose: *<-khu ~ -khus->* (AP):

(67) YAMPHU
akko Hedaŋna-beʔ peŋ-**ghu**ʔ
that Hedangna-LOC stay-**AP**
'Is he the guy who is [currently] staying in Hedangna?'

This is not the only agentive nominalization, however. A different nominalizer may be used if the agentive role is being portrayed as characteristic, not just a temporary situation.

(68) YAMPHU
na seʔ-**yaŋ**-ji
fish kill-**AGP**-NS
'fishermen'

This distinction of whether or not the agentive role is characteristic poses a real challenge for adaptation, as Limbu apparently does not make this division.

Yamphu can actually nominalize not only the agent of a verb (into the 'active participle'), but also the patient into a 'passive participle', certain instruments into an 'object participle', and locations into a 'locative' participle. This is somewhat exceptional in Kiranti, where most languages must resort to relative clause constructions or complementation in order to reference these other roles. This in itself is not an obstacle for adaptation into Yamphu, however. It simply means that text from Limbu will not result in locative participles being generated in Yamphu.

Greater challenges can be observed in nominalizations of finite verbs. In Limbu, the same <*-pa*> marker is utilized.

(69) LIMBU

anchige	thunʔ-ɛ-tch-u-ge-**bɛ**-n	thi:	kudzaphɛʔr-ɛ
we^{de}	drink-PT-dA-3P-e-**NOM**-DEF	beer	taste.bad-PT

'The millet beer we^{de} drank tasted bad.'

In such constructions, Yamphu uses either the 'factitive' marker <*-æ*> (FCT) or the 'plural nominalizer' marker <*-ha*> (PLNR), the choice of which is determined by number agreement.

(70) YAMPHU

am-mi	cabaŋ-æʔ	khi:-ghi:-tt-**æ**		mottitel-so	ha:-dis-e
your-GEN	guest-ERG	carry-bring.for-PF-**FCT**		kerosene-too	light-apply-IMP

'Light the kerosene which your guest has brought, too.'

(71) YAMPHU

mo-**ha**	eŋ-ʔiŋ-**ha**-reʔ	yuŋ-ma-ho	la:-ghiʔ-m-**æn**-ji	
that-**PLNR**	remain.NP.**PLNR**.only	put-INF-LCQ	take-bring.for-INF-**FCT**-NS	

'Take home for them only that which you put aside of what remains.'

Again we see a distinction (here based on number agreement) that Yamphu makes that is unmarked in Limbu. Computationally determining number agreement by reference to the syntax (for example, in (69) to *thi:* 'beer') is complex and sometimes impossible, such as when the head is not overt (e.g. *"the one(s) that we saw"* vs. *"the boy/boys that we saw"*).

We have barely scratched the surface of nominalization in Kiranti languages, and yet already we can observe that it presents a significant number of thorny obstacles for adaptation. Further research will be required in order to determine whether typological similarities can be utilized to surmount the structural differences.

9.3. Clausal Issues

9.3.1. Sequencing

Rutgers (1998:76) states that in Yamphu the normal means to express a sequence of events when relating a story is to use the 'sociative gerund' <*-nu ~ -nuŋ*> (SOC), the same marker that, when applied to nominals, is called the 'sociative case marker' (cf. Example (50)):

(72) YAMPHU

mo-baŋ-go	thapnam-beʔ	khæʔ-**nuŋ**	naŋkhi	to:s-i-ŋa,	naŋkhi
that.ELA.TH	forest.LOC	go.SOC	naŋkhi	dig.12PL.EXPS	naŋkhi

waham-jas-i-ŋa
boil.eat.12PL.EXPS

'[After] going into the forest, we dug naŋkhi roots and boiled and ate them.'

This marker attaches only to non-finite verbs, and somewhat surprisingly, it may even do so where the two verbs' subjects are not coreferential:

(73) YAMPHU

ham-p-te	ye:p-**nuŋ**	pho:to	khic-ba	læ:-tt.æʔ
where-LOC	stand-SOC	photo	make.pic-NOM	do-PF-FCT

'Where were [you] standing when he took the photo?'

This marker is, as we have discussed, cognate with Limbu's 'commitative' marker <*-nu*>. Indeed, a similar sequencing role can reportedly be played by this marker in Limbu narratives (Webster, p.c. 2006). However, the more typical sequencer is the suffix <*-aŋ*>, which can coordinate verbs (both finite and non-finite), adverbs, and clauses:

(74) LIMBU

khɔrɛːŋ	khɛːks-u-ŋ-**aŋ**	caŋ
bread	break.piece.off-3P-1sA-**and**	eat.1sA›3P

'I shall break off a piece of bread and eat it.'

Another semantically related but grammatically distinct role it plays is as a postposition on nominals, meaning *'also, too'*:

(75) LIMBU

aŋga	seːdzɔnwa-**ʔaŋ**	thiː-**ʔaŋ**	kɛrɛk	thuŋ-u-ŋ
I	millet.brandy-**too**	millet.beer-**too**	everything	drink-3P-1sA

'I also drink millet brandy, millet beer too and everything.'

In this role, it parallels Yamphu's 'inclusive focus' marker *<-so ~ -soŋ ~ -son->*, which may be glossed *'too'* or, especially where stringing together a depiction of a scene, *'and what is more'*.

(76) YAMPHU

beʔma-ma	pusæːt-thappa,	ya-**so**	pheːbhe.
big-ATNR	snake-big	face-**too**	wide_open

'It was a huge snake, and it had its mouth wide open, too.'

For adaptation purposes, where the Limbu *<-aŋ>* suffix appears on a nominal or on a finite verb, for the purposes of tranfer to Yamphu it seems the 'inclusive focus' marker is the most appropriate match. Where it appears on non-finite verbs, however, it would seem that the best option might be to apply Yamphu's sociative marker in its place, dropping the finite markers. This is definitely a problematic area.

Somewhat related to these sequencing issues are the notions of simultaneous action. Limbu's 'present gerund' *<-lə>* (prG) seems to be best paralleled by Yamphu's 'simultaneous gerund' *<-sæːʔ>* (SMG):

(77) LIMBU

luŋ-ʔoː	phɛdzaː-n	hasuk-**lɔ**	yutt-u-ŋ-loː!
stone-LOC	knife-DEF	be.sharp-**prG**	whet-3P-1sA-ASS

'I'm whetting this knife sharp against a stone!'

(78) YAMPHU

ap-pe-nuŋ	mo	pusæːʔ-mi	kha	iː-**sæːʔ**	iː-**sæːʔ**	ab-a-j-iŋ
come.RES.SOC	that	snake.GEN	word	say.**SMG**	say.**SMG**	come-PT-DU-EXPS

'We came, talking all the while about the snake.'

Note that in Limbu, both sequencing suffixes <*-lɔ*> and <*-aŋ*> are used in the construction of periphrastic tenses. (cf. 'present perfect' in section 5.4) Where such tenses are not periphrastic in some Kiranti languages, the tense is instead reflected with an appropriate functeme, and in such cases, the sequencing suffixes themselves are not passed through in the intermediate form.

9.3.2. Causal clauses

Both Limbu and Yamphu use a marker formally and semantically resembling the ergative-instrumental suffix to indicate a causal relationship between one clause and another subordinated clause. In Limbu, van Driem (1987:230) refers to it as 'the <*-ille*> subordinator' (SUB), while in Yamphu, Rutgers (1998:274) calls it the 'instrumental gerund' <*-æʔ*> (INS).

(79) LIMBU

yaːmbɔk	cok-mɛ-**lle**	naːs-aŋ	khips-aŋ
work	do-INF-**SUB**	tire-1sPS.PT	jingle-1sPS.PT

'I have gotten tired from doing the work.'

(80) YAMPHU

'sip-pe-peː-tt-æn-de?'	kaː-nuŋ	piss-a	tham-so	thaps.a,	sapthaŋ-m-**æʔ**.
fall-RES-PF-FCT-ISF	cry-SOC	RUN-PT	fall-TOO	fall-PT	rejoice-INF-**INS**

'Have we caught one?' he cried and ran, falling from excitement.

One difference, however, is that in Limbu, the marking may occur on a finite verb.

(81) LIMBU

hɛkkeː	kɛ-baːtt-u-m-**ille**	a-niŋ	lɛʔ	lɛʔ!
like.that	2-speak-3p-pA-**SUB**	1-ire	unleash	unleash

'If you[p] are going to talk that way, I'll get fed up!'

In the Yamphu corpus, it appears that no form of the ergative-instrumental marker <*-æʔ*> can be suffixed to finite verbs. (Where affixed to nominals, it is

labeled ERG, and where attached to infinitive verbs, it is labeled INS.) Thus, it cannot be used where the Limbu SUB marker is suffixed to a finite verb. In such contexts, for the purposes of transfer we map it to Yamphu's 'logical consequence' marker <-*hoŋ ~ -ho ~ -hon*-> (LCQ), which may attach to finite verbs.

(82) YAMPHU

i-doʔ-noʔ	ma:d-a-**hoŋ**	kaniŋ-æʔ	i-doʔ	akkraŋ-beʔ
this-like-EXF	not.be-PT-**LCQ**	we^{pe}-ERG	like-this	shoulder-LOC

paŋ-ʔænd-u-ŋ-ma
hang-put.down-›3-EXAG-12NS

'Since there weren't things like [needles], we hung [the cloth] over our shoulders like this.'

The function of this 'logical consequence' marker is to indicate that relationship with the subordinated clause is one of cause, sequence, or general dependency (Rutgers 1998:274,312). It thus seems well-suited to convey the semantics of Limbu's finite subordinated clause structure.

9.3.3. Assertive/Emphatic particle

Limbu has a fairly high-frequency clause-final particle[17] *lo:/ro:* that van Driem (1987:242) calls the 'assertive particle' (ASS), describing it as making "an appeal… to the listener to pay attention and heed the *implications* of what is being said" (emphasis in the original).

(83) LIMBU

kɛ-gen-**lo:!**
2-stumble.and.fall-**ASS**
'You^{s}'ll stumble and fall if you don't watch out.'

Yamphu, too, has an 'assertive' marker <-ye:> (ASS). "An assertive clause has roughly the same communicative effect as the phrase 'Hey, I tell you…', but simultaneously expresses an emphatic appeal toward the hearer to accept or acknowledge what is said" (Rutgers 1998:305). While this seems to be some semantic overlap with Limbu's 'assertive', perhaps even better semantic

[17] It does not appear in the list of high-frequency words in section 4.2, as orthographic standardization efforts have called for it to be written as a bound morpheme.

alignment may be Yamphu's 'exhortative suffix' <-*ti*> (EXH), which "expresses an appeal for the hearer to acknowledge the purport of the message conveyed." Unlike Limbu's 'assertive', however, Yamphu's 'exhortative' is not necessarily clause-final, as it may attach to other constituents, by which it marks them as the focal argument of the clause. Further complicating the matter, when this suffix is attached to the predicate verb (that is, clause-finally, where its function is akin to Limbu's 'assertive'), if the verb is finite, a factitive marker (the nominalizer discussed on page 113) must also be present:

(84) YAMPHU

i-beʔ	yaʔmi	ceŋ-ʔitt-u-ji-ro-**en-di**
this.LOC	person	cut.PF.›3.3NS.REP.**FCT.EXH**

'I hear that they've killed somebody here.'

It is thus with a great deal of uncertainty that we might tentatively propose to transfer Limbu's 'assertive particle' to Yamphu's 'exhortative suffix', inserting an additional 'factitive' marker when the verb is finite. This would definitely require thorough testing with Yamphu speakers. If unacceptable, the best remaining option might be to completely ignore the Limbu 'assertive particle' for the purposes of transfer to Yamphu, much like the deletion strategy discussed on in section 2.3 (specifically page 21).

9.3.4. Reported speech

One other high-frequency feature present in virtually all languages of the region (including Nepali with its particle *re*) is that of marking a clause's predication as being second-hand, as a hearsay evidential. In Limbu, the 'reported speech particle' is <*mu*> (REP):

(85) LIMBU

mɛ-be:k-pa	**mu**
nsAS-go-IPF	**REP**

'They say they're going.' / 'I hear they're going.'

In Yamphu, the 'reportative suffix' is <-*lo*> (REP):

(86) YAMPHU

e, dobha:n-be?-mu khæ:.tta.**ro**?
oh confluence-LOC-DWN go-PF-**REP**
'Oh, did he say he went to the confluence?'

This functional and structural commonality makes for smooth adaptation here.

9.4. Examination of Parallel Texts

Having examined the broad categories to assess the degree to which Limbu and Yamphu "care about the same kind of stuff", the next step in the assessment would be to compare a Limbu text with a fairly literal Yamphu translation of it (performed by a human translator), to consider whether the transformations required are feasible. Unfortunately, a short-coming of the selection Yamphu as a target language now becomes evident: Yamphu speakers outside the language area are few and far between, and due to the civil war in Nepal, it has not be feasible for me to visit the language area. Indeed, I have never personally met a Yamphu speaker. Limbu-to-Yamphu adaptation can only ever be declared successful if the results are comprehensible to a Yamphu speaker. In the meanwhile, however, another textual strategy may help us to assess feasibility and to pilot strategies for adaptation: No Yamphu translation of a Limbu text is known to exist, and the Yamphu texts that are available (in Rutger's grammar), are transcriptions of speech, which, at that, is essentially colloquial in nature. However, taking a portion of one of these interlinearized Yamphu texts[18], Limbu translators working with Webster attempted to produce the closest Limbu equivalent. These Limbu translators were not Yamphu speakers, however, and so the texts are probably not as closely paralleled as might be possible. In some places, alternative Limbu renderings were provided (labeled 'Alt').

[18] Extracted from the text "Buffalo Hunt" (Rutgers 1998:342)

Clause by clause, we shall look over these parallel texts, and consider a few implications for automated transfer. (The upper pair, labeled '**Y**', represents the Yamphu source, while the lower pair labeled '**L**' represents the Limbu translation.)

(87) 'A loud noise resounded through the entire jungle.'

Y	jaŋgal	bonpala.itthuk		ikko	awa:j	ka:sa.
	jungle	forest.entire		one	noise	cry.PT
L	jaŋgal	kʰarak	yɔmba	ikla.dʰik.lɛ		lo:kkʰ.u
	jungle	entire	big	sound.one.ERG		resound.3sPT

Here, the Yamphu noun *bonpala* (which may be related to Nepali *bən 'forest'*) is glossed as 'forest'. In other Yamphu texts, a different word, *nambhuŋla*, is given the same gloss, so there are probably distinctions (e.g. based on vegetation, altitude, or terrain) in the types of places referred to by these expressions. The issue for adaptation is to choose the most appropriate Yamphu form. *jaŋgal* is, of course, a wide-spread borrowing. That Limbu uses simply "jungle" where Yamphu can compound "jungle-forest"—or even use the other form mentioned—is probably not a big obstacle to comprehension. In this case of *multiple lexical alternatives*, it would probably work to simply transfer only *jaŋgal* and leave the other possibilities aside.

The Yamphu word *itthuk 'entire'* can be an adjective or an adverb, or (as in this case) a postposition that corresponds to the Limbu adjective/postposition *kʰarak 'entire/throughout'*. This is *a well-matched pair*, and so the Limbu form should be able to transfer to the Yamphu form quite consistently.

The next two Yamphu words (*ikko awa:j 'one noise'*) are translated with a single Limbu word with *constituents in the reverse order*, but this does not in itself pose a difficult obstacle for transfer. More significant is the *missing ergative marker*: The ergative/instrumental marker (ERG) required in Limbu is not present in the Yamphu. Rutgers lists the variety of overlapping contexts in Yamphu in which the ergative marker would be required, and this would seem to fit those. Thus, in transfer to Yamphu from Limbu, it should not be

necessary to formulate a rule to drop the ergative marker here. Even if its presence is unnecessary here, retaining it should still do no real damage to comprehensibility. Perhaps its absence here is merely an indication of its optionality in colloquial speech.

The final word in (87) to examine is the verb, in both languages an intransitive verb with past tense and third person singular agreement (with the noise/sound). However, as we will observe in (91) and (94), the *semantic range* of the Yamphu verb *ka:sa 'rang out / cried out / called'* is not a perfect match with the Limbu verb *lo:kkʰu 'resounded / sounded / rang out'*. Indeed, issues of semantic misalignment are extremely complex to address computationally.

Consider now the next sentence of our parallel text:

(88) 'I look. It was a large snake.'

Y	khaŋ.ʔin.uŋ.æ		pusæ:t.thappa	læ:tta
	see.NP.1›3.FCT		snake.big	be.PT
L	ɔmɔtt.u.ŋ.[ŋillɛ].	yɔmba	ɔse:k.kin	wɔy.ɛ
	look.3s.1sA.[TEMP]	big	snake.DEF	be-exis.PT

Here we see *mismatched nominalization*: Yamphu's factitive nominalizer FCT is used in a context where Limbu would not use a nominalizer. Indeed, use of Limbu's nominalizer here was rejected by the Limbu translators. (It may seem that here it corresponds to Limbu's temporal marker (TEMP), but actually, the temporal marker apparently has quite a different function. It is optional here, but the Limbu translators felt it would be much more natural to include it as a means of joining the verb with the rest of the sentence.)

Both Limbu and Yamphu have a both a regular adjective *'big'* (to be exemplified for Yamphu in (89)) as well as a suffixal *'big'* modifier (to be exemplified for Limbu in (94)). It is unclear to me what governs the choice between regular adjective versus the suffixal form.

(89) 'It was a huge snake, and it had its mouth wide open, too.'

Y		be?ma.ma	pusæ:t.thappa,	ya.so	phe:bhe.
		big.ATNR	snake.big	face.too	wide_open
L	sarik	yɔmba	ɔse:k.kin.	ku.mura	pʰaks.u.ba
	very	big	snake.DEF	3poss.mouth	open.3s.NOM

In the Yamphu text, there is no verb here (*phe:bhe* is an adjective), perhaps because in this colloquial speech, this description is being added as an afterthought to the prior clause. The Limbu verb here can be used of a book, or something folded open. In Yamphu, it is the snake's "face" that is open, whereas this is not possible in Limbu, where it must be the snake's "mouth". This, then, would seem to be an *idiomatic difference* that adaptation must contend with.

(90) 'Then Kancha also came.'

Y	mo.ba	kancha.so	less.a.
	that.ELA	Kancha.too	come.PT
L	hɛkkyaŋ	kancʰa.aŋ	tyɛ
	then/and	Kanch.also/and	come(3sPT)

On page 102 we discussed adapting Limbu's *hɛkkyaŋ* to Yamphu's *moba*. That they are matched in the above sentence exemplifies such an alignment. This sentence also exemplifies the paralleling of Yamphu's 'inclusive focus' marker <*-so*> with Limbu's multi-functional <*-aŋ*> suffix where it appears on nominals (cf. p. 115).

(91) 'Upon the snake having suddenly made this noise, I fell back from fright and unexpectedly landed, hanging across Kancha's shoulder.'

Y		pusæ?	swa:kto?	ka:tt.æm.be?		ka	cain.ghæ?.nun
		snake	suddenly	cry.PF.FCT.LOC		I	get_a_fright.go.SOC
L	(kʰɛn)	ɔse:k.killɛ	hɔkcɔgɔt	se:kt.ɛ.llɛ			kis.aŋ.ŋaŋ
	(that)	snake.ERG	suddenly	hiss.PT.TEMP			afraid.1sP.and
Alt				se:k	lɔ?r.ɛ.llɛ		
				hissingly	said.PT.TEMP		

Y	thaps.iŋ.æm.be?			khw.e?e	akkr̂aŋ.be?	paŋ.drus.iŋ,	kanch.æ?æm.be?	
	fall.1S.FCT.LOC			s/he.POS	shoulder.LOC	hang.CEX.1S	kancha.POS.LOC	
L	cillɛk	lɛkkʰ.aŋ.ŋillɛ		kancʰa.rɛ	ku.bʰɔktaŋ.ŋo			tʰy.aŋ
	backwards	fell-back.1sP.TEMP		Kancha.GEN	3poss.shoulder.LOC			fall.1sPS
Alt			kancʰa.rɛ	ku.bʰɔktaŋ.ŋo	cillɛk	lɛkkʰ.aŋ	tʰy.aŋ	
			Kancha.GEN	3poss.shoulder.LOC	backwards	fell-back.1sP	fall.1sPS	

This is the most complex sentence in the text, the speaker's emotional state on encountering the snake seemingly reflected in a scrambling for words as he describes that instant, the final word *kanch.æʔæm.beʔ* (kancha.POS.LOC) *"on Kancha's"* clarifying the referent of the earlier genitive pronoun in *khw.eʔe akkɦaŋ.beʔ* (s/he.POS shoulder.LOC) *"on his shoulder"*. The Limbu translators, however, preferred to name Kancha directly in place of using a pronoun that gets belatedly clarified, a strategy which ought to be fine in Yamphu, too. Note also, here, that Yamphu has used its POS genitive—not GEN—to correspond to the Limbu genitive (cf. p. 96).

Again we see a semantic misalignment between Yamphu's apparently wider-purpose *'cry out'* verb and Limbu's *'hiss'*. Apparently, in Yamphu semantics, snakes—like men—can *'cry out'*, while in Limbu semantics, snakes cannot *'cry out'* and are instead preferred to *'hissingly speak'*. For adaptation from Limbu, should we transform *'hissingly speak'* to Yamphu's *'cry out'*? Well, Rutger's lexicon does provide us with an onomatopoeic Yamphu adverb *cwæːŋdoʔ*, described as *"with a sizzling or hissing noise, with a sizzle, with a hiss, as when water is heating up or when a drop of water evaporates from a hot surface."* If this parallels the function of Limbu's adverb *seːk 'hissingly',* perhaps the *'hissingly speak'* idiom will carry over into Yamphu.

In the Yamphu text, the backwards receding motion is marked on the *'get-a-fright'* verb by means of the 'general receding motion auxiliary' otherwise glossed *'go'* (cf. Rutgers 1998:145). As can be seen, this backwards motion is expressed by other means in Limbu. Indeed, the preferred word-ordering in Limbu puts the verb and backwards motion clause-finally.

That the event described was unexpected is indicated by the 'contrary to expectation' marker (**CEX**) which actually has two forms, determined by transitivity: <*-trus*> on intransitive verbs, and <*-trid*> on transitive ones.

(92) Yamphu

superwaiser.so leŋ.ʔa.**dru:**.tta
supervisor.too come.PURP.**CEX**.PF
"The supervisor also unexpectedly turned up."
(Rutgers 1998:192)

This is distinct from mirativity, but apparently the semantics of unexpectedness predispose it to co-occur with the mirative marker (cf. both markers in Kham, Watters 2002:296).

(93) 'Then Kancha grabbed me. He also caught sight of the snake.'

Y	mo.ba	kanch.æʔ	ra:b.a.	kho.es.so			khaŋ.dog.u.
	that.ELA	Kancha.ERG	seize.PT	s/he.ERG.too			see.find.›3
L	hɛkkyaŋ	kanchʰa.rɛ	hɛpt.aŋ.	khunɛʔ.aŋ	ɔse:k.kin		ni:ss.u
	then	Kancha.ERG	grab.1sP	3prn.also	snake.DEF		see.3s
Alt				kʰɛl.lɛ.aŋ		hɛp lɔrik	ni:ss.u
				3sdem.ERG.also		suddenly	see.3s

It is problematic for adaptation that the Yamphu verb *<ra:p-> 'seize'* does not match the semantics of the Limbu verb root *<hɛpt-> 'grab'* in all contexts. The Limbu verb apparently involves the notion of putting arms around something in hugging fashion. The Yamphu verb, on the other hand, may be used in the contexts of catching chickens or tadpoles.

On the final verb, Yamphu uses the 'auxiliary of opportunity' (formally identical to the independent transitive verb *<-tɔg-> 'find'*) to give the sense '*he got to see'* (Rutgers 1998:185). This option is not open to Limbu, but the Limbu translators did not like to simply say *'he saw'*, and so added the adverb *hɛp lɔrik 'suddenly seeing'*.

(94) '"Wow, what a really huge one," he cried.'

Y	'abhui	abhui	indo.dhappa.de,'				ka:s.a.
	EXCL	EXCL	like_what.big.ISF				cry.PT
L	abhui	abhui	akkʰɛn.gyappa.ni	be	lɔʔrik		pʰikt.ɛ
	EXCL	EXCL	how_much.size-suff.EMPH	go	one(emph)		cry-out.PT
Alt	ammwi	ammwi					
	EXCL	EXCL					

A parallel is suggested here between Yamphu's 'insistive focus' marker *<-te>* (ISF) (Rutgers 1998:287) and Limbu's emphatic marker *<-ni>* (EMPH).

However, further investigation would be required in order to confirm this.

Apparently, in this context, the emphatic form of the numeral *'one'* serves as an additional means of marking the clause as emphatic.

(95) 'Both of us started to cry out.'

Y	kajiŋ	nip.paŋ.no?	ka:.bug.a.jiŋ	
	we^{de}	two.UN.EXF	cry.start.PT.du1S	
L	anchi?ge	nɛpma	p^{h}ik.ma	he:kt.u.si
	we^{de}	both(dual)	cry-out.INF	start.3s.np

Here Yamphu's 'inceptive auxiliary' <*-pug*> (start) exemplifies how Yamphu auxiliaries are attached directly to the main verb's root. In transfer from Limbu, the intervening infinitive marker <*-ma*> (INF) is dropped and the words merged.

This examination of parallel texts has thus served to highlight a number of issues that are not highlighted by a comparison of the major typological categories, perhaps most significantly the issues of semantic range and idiomatic use.

10. CONCLUSIONS

10.1. Assessments

The overview of machine translation strategies has enabled me to describe the current study's strategies in terms of that larger framework, and to acknowledge our many debts to earlier pioneers. I have also been challenged to consider ways in which recent innovations in statistically-based systems may be utilized in the resource-scarce situations faced by minority languages.

The examination of issues of word frequency in Limbu has demonstrated how Limbu's highly-complex affixation results in a dramatically less effective situation for translation-memory strategies. The strategies utilized instead can be classified as an interlingua-hybrid of a transfer-based system, where that "interlingua" is specifically a Kiranti interlingua. This interlingua embodied the philosophy that what really matters for adaptation is that languages "care about the same kind of stuff," even if they encode it in structurally diverse means. I have demonstrated that even where it may be impossible to reliably map morphemes (such as the Kiranti agreement markers) directly from one language to another, a function-oriented approach could extend the reach of transfer-based adaptation there. We can thus make Limbu-to-Yamphu adaptation reach further (especially with finite verbs), but still unresolved are the questions of whether that extended reach is itself "far enough."

This raises the question of the suitability of the selection of source and target languages. In Kiranti typological comparisons, Limbu often seems to stand out as somewhat exceptional. Where it is making distinctions that other Kiranti languages are not (such as in certain agreement patterns), this is no problem. However, where Limbu uses a single general structure that is more specialized in the target language (such as nominalizers), adaptation gets much more difficult. Yamphu, too, has demonstrated some challenges

seemingly unique within Kiranti, such as its split genitive system, the type of distinction that is awkward to find in a target language. It may be that this type of phenomenon would be best served by the incorporation of a corpus-based strategy, though that is no small matter. Perhaps these same strategies applied to closer languages would be more fruitful.

I have also introduced a strategy for disambiguation that incorporates a feedback process, so that successive target languages can take advantage of the manual disambiguation efforts performed earlier in other target languages. Indeed, as the *AnglaBharti* project, like ours, involves a multi-language target strategy, I believe that our disambiguation feedback strategy could prove fruitful for them.

A typological comparison of Limbu and Yamphu demonstrates a number of striking similarities and a number of striking differences. What these really boil down to in terms of comprehensibility issues is difficult to gauge, however. Some degree of awkwardness is acceptable if reasonable comprehensibility is achieved. For example, if a program that adapted Kham into English produced the awkward sentence, *"He cooked the you-brought-them chickens,"* an English-speaking post-editor may be able to turn this into a more natural relative clause: *"He cooked the chickens that you brought."* It is also likely that the post-editor would soon learn to recognize the types of awkward structures that result. For example, the Yamphu post-editor would soon recognize that the adaptation process often uses the less-fitting genitive. On the other hand, when faced with the phrase "retain an actuary and patrol" in (1), this may be beyond the powers of an English-speaking post-editor to fathom the intended meaning. There may be equally-baffling results in Yamphu, where we have certainly not invested the millions of dollars that, for all that, still only resulted in that puzzling Spanish-to-English translation. In such cases, it seems that a Yamphu post-editor must then either seek the assistance of a Limbu-speaker, or, if the source material

is also available in Nepali, refer to the Nepali version for the meaning, with the adaptation results serving only as a window on how the idea was cast in Limbu.

10.2. Areas for Future Investigation

Clearly, the next step would require the involvement of a Yamphu speaker, ideally one who also spoke Limbu, and could thus construct a fairly literal Limbu-to-Yamphu translation to serve as an example for machine translation. Even a Yamphu speaker who did not speak Limbu would be able to provide the vital assessment of how well (or at least how comprehensibly) the output of our proposed alignments and rules communicates. Further investigation would also be required in many areas of Kiranti typology not yet sufficiently examined. These areas include the use of auxiliary verbs, other tense and periphrastic constructions, issues of focus, a variety of postpositions, and the incredible variety of functions of nominalization.

Moreover, the typological comparison needs to be widened across more languages. Several closely-related Kiranti languages lack even the most basic descriptive grammar. Even in languages so well described as Limbu and Yamphu, sometimes half a dozen example sentences of the function of a given morpheme may still be insufficient to truly identify whether the respective morphemes are in alignment, or whether the authors approached the same semantic phenomenon like the blind men of the fable, who grasped different parts of the elephant, and so each described quite differently what he had encountered.

It is only as we gain a real understanding of the typological unity and differences that exist in Kiranti languages that we will gain an accurate picture of the potential there for automated adaptation to benefit the Kiranti language communities.

REFERENCES

Al-Onaizan, Y., U. Germann, U. Hermjakob, K. Knight, P. Koehn, D. Marcu, and K. Yamada. 2003. 'Translation with Scarce Bilingual Resources", *Machine Translation.*

Bar Hillel, Y. 1959. *Report on the state of machine translation in the United States and Great Britain.* Technical report, 15 February 1959. Jerusalem: Hebrew University.

Bickel, Balthasar. 1999. Nominalization and focus constructions in some Kiranti languages. In *Topics in Nepalese linguistics*, eds. Yogendra P. Yadava & Warren W. Glover, 271-296. Kathmandu: Royal Nepal Academy.

Black, A. and C. Black. 2005. *A conceptual introduction to morphological parsing using AMPLE.* Grand Forks, ND: SIL.

Bradley, David. 2002. The subgrouping of Tibeto-Burman. In *Medieval Tibeto-Burman languages. piats 2000: Tibetan studies: Proceedings of the ninth seminar of the international association for Tibetan studies.*, ed. LEIDEN 2000. Beckwith, Christopher I., 73-112. Leiden, Netherlands; Netherlands: Brill.

Brown, P., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer: 1993. 'The Mathematics of Statistical Machine Translation: Parameter Estimation', *Computational Linguistics* **19**, 263–311.

Charniak, E., K. Knight, and K. Yamada. 2003. *Syntax-based Language Models for Machine Translation.* Proc. MT Summit IX, 2003.

Cunningham, Robert. 2006. *Genitive is not always possessive.* Available at: http://alt-usage-english.org/genitive_and_possessive.html

Czuba, K., T. Mitamura and E. Nyberg. 1998. *Can Practical Interlinguas Be Used for Difficult Analysis Problems?* Proceedings of AMTA-98 Interlingua Workshop. Available at: www.lti.cs.cmu.edu/Research/Kant/PDF/amta98-irw.pdf

DeLancey, Scott. 1981. The category of direction in Tibeto-Burman. *Linguistics of the Tibeto-Burman Area* 6.1:83-101.

DeLancey, Scott. 1997. Mirativity: The grammatical marking of unexpected information. *Linguistic Typology*, *1*, 33-52.

Doornenbal, Marius. 2004. *Limbu morphological parsing*. (ms) Kathmandu.

Driem, George van. 1987. *A Grammar of Limbu*. Berlin: Mouton.

Driem, George van. 1999. *The Limbu verb revisited*. In *Topics in Nepalese linguistics*, eds. Yogendra P. Yadava & Warren W. Glover, 209-230. Kathmandu: Royal Nepal Academy.

Driem, George van. 2001. *Languages of the Himalayas: An Ethnolinguistic Handbook of the Greater Himalayan Region*. Leiden, Brill.

Ebert, Karen H. 1994. *The Structure of Kiranti Languages: comparative grammar and texts*. Arbeiten des Seminars für Allgemeine Sprachwissenschaft, Nr. 13. Zürich: Universität Zürich.

Givón, T. 1971. Historical syntax and synchronic morphology: an archaeologist's field trip. *Chicago Linguistics Society* 7:394:415.

Gordon, Raymond G., Jr. (ed.) 2005. *Ethnologue: languages of the world,* Fifteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

Grimes, Barbara (ed.). 1996. *Ethnologue*: *languages of the world*, thirteenth edition. Dallas: Summer Institute of Linguistics.

Harley, Heidi and Elizabeth Ritter. 2002. Structuring the bundle: A universal morphosyntactic feature geometry In: Simon, Horst J. and Heike Wiese (eds.), *Pronouns – Grammar and Representation*. Philadelphia: J. Benjamins (pp. 23–39). Available at: dingo.sbs.arizona.edu/~hharley/PDFs/HarleyRitterMarburg2000.pdf

Hart, Robbie. 2004. *The Final Frontier: Spatial Terminology in Kiranti*. Swarthmore, PA. Available at: http://www.swarthmore.edu/SocSci/Linguistics/papers/2004/hart.pdf

Homola, Petr, and Vladislav Kuboň. 2005. *A Machine Translation System into a Minority Language*. Paper presented at RANLP Workshop 2005 (Borovets, Bulgaria, September 24, 2005). Available at: nats-www.informatik.uni-hamburg.de/view/RANLPMT2005/WebHome

Hong, Munpyo, and Oliver Streiter. 1999. *Overcoming the language barriers in the Web: The UNL-Approach*. In GLDV 1999. Available at: http://www.iai.uni-sb.de/docs/unl_gldv.pdf

Hutchins, W. John. 1995. 'Machine translation: a brief history'. In: *Concise history of the language sciences: from the Sumerians to the cognitivitsts.* Edited by E.F.K. Koerner and R.E. Asher. Oxford: Pergamon Press, 1995. pp. 431-445.

Hutchins, W. John. 2003. 'Machine translation: general overview'. In: *The Oxford Handbook of Computational Linguistics*. Edited by Ruslan Mitkov (Oxford: University Press, 2003), pp. 501-511.

Jurafsky, D. and Martin, J. H. 2000. *Speech and Language Processing.* Prentice-Hall.

Knight, K. 1999. *A Statistical MT Tutorial Workbook, Prepared in Connection with the JHU Summer Workshop*. Technical report, USC/ISI, Los Angeles, CA. Available at www.isi.edu/natural-language/mt/wkbk.rtf.

Koehn, P. and Knight, K. 2003. Feature-rich statistical translation of noun phrases. In Proceedings of the 41st Annual Meeting on Association For Computational Linguistics - Volume 1 (Sapporo, Japan, July 07 - 12, 2003). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 311-318.

Lavie, A., Probst, K., Peterson, E., Vogel, S., Levin, L., Font-Llitjos, A., and Carbonell, J. 2004. *Trainable Transfer-based Machine Translation Approach for Languages with Limited Resources.* In Proceedings of the 9th European Association for Machine Translation (EAMT) Workshop (Malta, 26-27 April 2004). University of Malta.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass: The MIT Press.

Nießen, S. and Ney, H. 2004. Statistical Machine Translation with Scarce Resources Using Morpho-syntactic Information. Comput. Linguist. 30, 2 (Jun. 2004), 181-204. Available at: http://dx.doi.org/10.1162/089120104323093285

Opgenort, Jean Robert. 2005. *A grammar of Jero: with a historical comparative study of the Kiranti languages*: Languages of the Greater Himalayan Region, vol. 5/3. Leiden; Boston: Brill.

Rao, Durgesh. 2001. *Machine Translation in India: A Brief Survey*. Available at: www.elda.org/en/proj/scalla/SCALLA2001/SCALLA2001Rao.pdf

Rutgers, Roland. 1998. *Yamphu: Grammar, Texts & Lexicon.* Leiden: Research School CNWS.

Sato, S. and M. Nagao. 1990. Towards Memory-based Translation. In *Proceedings of COLING-90*. 247-252.

Silberman, Steve. 2000. Talking to Strangers. *Wired*, 8.05. The Condé Nast Publications Inc.

Sinha, R.M.K. and A. Jain 2003. *AnglaHindi: An English to Hindi Machine-Aided Translation System*. MT Summit IX: Proceedings of the Ninth Machine Translation Summit, New Orleans, USA, September 23-27, 2003. Available at: http://www.amtaweb.org/summit/MTSummit/FinalPapers/36-sinha-final.pdf

Slocum, J. 1985. *A survey of machine translation: its history, current status, and future prospects.* Comput. Linguist. 11, 1 (Jan. 1985). pp. 1-17.

Toba, Sueyoshi. 1984. *Khaling*. Tokyo: Tokyo University.

Turin, Mark. 1998. The Thangmi verbal agreement system and the Kiranti connection. *Bulletin of the School of Oriental and African Languages* 61, (3) (Oct). pp. 476-491.

Watters, David. 1988. *CADA: The Kham Experiment.* Grand Forks, ND: Summer Institute of Linguistics.

Watters, David E. 2002. *A grammar of Kham*. Cambridge Grammatical Descriptions, ed. by R. M. W. Dixon and Keren Rice. Cambridge UK: Cambridge University Press.

Watters, David. 2006. *Nominalization in Himalayan languages.* (ms.) To appear in *An Encyclopedia of Nepal's Languages*. Kirtipur: Tribhuvan University.

Watters, David and Dan Raj Regmi. 2005. *Bhujel "Direct–Inverse".* Paper presented at the 26th Annual Conference of the Linguistic Society of Nepal, Kathmandu, Nepal.

Weaver, W. 1949. 'Translation'. Repr. in: Locke, W.N. and Booth, A.D. (eds.) *Machine translation of languages: fourteen essays.* Cambridge, Mass.: Technology Press of the Massachusetts Institute of Technology, 1955.

Webster, Jeffrey. 2000. *Automated analysis of Limbu.* (ms.) Kirtipur: Center for Nepal and Asian Studies.

Weidert, A. and B. Subba. 1985. *Concise Limbu grammar and dictionary*, Amsterdam: Lobster Publications.

Yamada, K. and Knight, K. 2001. *A syntax-based statistical translation model.* In Proceedings of the 39th Annual Meeting on Association For Computational Linguistics (Toulouse, France, July 06 - 11, 2001). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ. pp. 523-530.

APPENDIX A

In the column on the left is the original Spanish-language news article discussed in Section 2 (as found at http://www.cronica.com.mx/nota.php?id_nota=259873).
In the column on the right is a human translation of the article.

Habitantes de Santa Catarina en Tláhuac retienen a actuario y patrulla	**Inhabitants of Santa Catarina in Tlahuac hold lawyer and patrol car**
Miercoles 6 de Septiembre de 2006 | Hora de publicación: 09:46	*Wednesday 6 September 2006| Time of publication: 09:46*
Habitantes del pueblo de Santa Catarina, delegación Tláhuac, retuvieron desde temprana hora una patrulla y a un actuario que iba a efectuar un desalojo, por lo que exigen la presencia de autoridades.	From an early hour, inhabitants of the town of Santa Catarina, in the Tláhuac delegation, held a patrol car and a lawyer who was going to carry out an eviction, which was the reason why they demand the presence of the authorities.
La Secretaría de Seguridad Pública (SSP) del Distrito Federal informó que al lugar ya se dirige el subsecretario Gabriel Regino para hablar con los inconformes y tratar que la situación se normalice.	The Federal District Public Safety Secretariat (SSP) announced that the undersecretary, Gabriel Regino, is already heading to the site to speak with the complainants/demonstrators and to try to get the situation normalized.
Esta mañana en la zona conocida como La Cruz, del pueblo de Santa Catarina, se presentó un actuario para realizar un desalojo y al no permitírselo alrededor de cien personas se pidió la presencia de la fuerza pública.	This morning in the zone known as "La Cruz" of the town of Santa Catarina, a lawyer presented himself to make an eviction, and when around one hundred people gathered to prevent the eviction, he requested the presence of the police.
Por esta razón arribaron elementos del cuerpo de granaderos y uniformados; sin embargo, los habitantes se apoderaron de la patrulla AC019 y detuvieron al actuario que iba a cumplir su trabajo.	For this reason, uniformed police and riot police arrived; however the inhabitants seized the patrol car AC019 and held the lawyer who was trying to do his job.
Ante esa situación se solicitó la presencia de más granaderos y la del subsecretario de Seguridad Pública capitalina, Gabriel Regino, quien ya se dirige a la zona para dialogar con los inconformes.	In light of the situation, more riot police and the presence of the capital's undersecretary of Public Security, Gabriel Regino, were requested. The undersecretary is already heading to the zone to engage in a dialog with the complainants.