САХА РУССКИЙ КОМПЬЮТЕРНЫЙ ПЕРЕВОДЧИК

Мигалкин Василий Васильевич
Физико-технический институт при Якутском государственном университете,
677006, г.Якутск, ГСП, ул.Кулаковского 48
e-mail: mjurju@mail.ru

TO THE PROJECT SAKHA – RUSSIAN COMPUTER TRANSLATOR

Migalkin Vasily Vasilevich
Physical-Technical institute of the Yakut State University,
677006, Yakutsk, Kulakovskogo 48
e-mail: mjurju@mail.ru

The present work will be the first attempt of semantic translation in Russia from one of the Turkic languages to Russian, which is of itself represents a great interest for philology science and as well as for education process.

Usage of benchmark language in the program of Russian language (possessing minimal number of homonyms and phrasal constructions) gives us belief in further translations from a fully synthetic language (Turkic) by means of Russian to fully synthetic languages (Turkic). The term – fully synthetic in this case is conditional and indicates the availability of Turkic language to use 3000 to 5000 affixes virtually to all words, which is impossible in Russian.

Presence of great number of homonyms, imitative words, pair and phraseology words, absolute difference in grammatical structure of sentences, many incentive word forms in Yakut language do not allow to use word – by – word translation method of completed idea. (Yakut language in complete decomposition represents more than 45 million words though the number of root words is only about 15000). The only method that can be used here is semantic one. Thus, there is a task to encode meaning of the sentence in the form of invariant mathematical formulae (for concrete meaning) regardless of the method of presenting the sentence.

Язык саха (якутский) является одним из древнейших живых тюркских языков прошедших своеобразный путь развития в условиях почти полной изоляции от других тюркских языков. Из-за своеобразного развития, язык саха в настоящее время стоит особняком в семье тюркских языков. Достаточно напомнить тот факт, что если носители других тюркских языков (за исключением чувашей) в той или иной мере могут понять друг друга, то это совершенно невозможно между якутами с одной стороны и

представителями других тюркских народов – с другой стороны. Именно такое развитие языка саха вызвало особый интерес к его изучению со стороны тюркологов с давних пор. Можно перечислить имена выдающихся исследователей первой половины XIX века В. Шотта, О. Бётлингка, В. Радлова, В. Ястремского, Э. Пекарского. В знаменитых экспедициях императорской Академии Наук возглавляемым академиком О.Н. Бётлингком, в достаточной мере исследовались фонетика, морфология и лексика языка саха. Наиболее полное исследование языка саха проводили политические ссыльные И.А. Худяков, С.В. Ястремский, Э.К. Пекарский, В.М. Ионов, труды которых востребованы и поныне. [1]

В чём основное резкое различие языка саха от славянской группы языков (по определению филологов - «синтетического русского языка», например). Конечно, по сравнению с аналитической романской группой языков, русский язык имеет значительно богатый однозначный лексический запас, наличие приставок, суффиксов и аффиксов придают к словам более тонкие оттенки значений слов, кратность действия, местоимение и так далее. Тем не менее, русский язык по сравнению с языком саха является всё-таки, полусинтетическим языком. В русском языке нет возможности указать местоимение к существительным, а также к глаголам в прошедшем времени с помощью аффиксов, сравнение объектов или движений производится исключительно аналитически, нет единого правила перехода глагола из несовершённой формы в совершённую форму, модальность в основном выражается аналитически (в языке саха от обязательно должного до обязательно не должного выражается синтетически в виде аффиксов в девяти градациях плюс два аффикса для выражения «что будет плохого, если ...» и «что будет положительного, если не ...» для любых глаголов). В русском языке очень мало производных глаголов (пила – пилить), а в языке саха с любого имени существительного или имени прилагательного можно образовать глагол (например: москвалаабатах – оказывается он не смог поехать в Москву или илимниэм – да, я собираюсь рыбачить сетями или ещё кыһыллыам – я покрашу в красный цвет) кроме-того в русском языке нет возможности выразить синтетически побуждение одного лица над другим (например: бултатыам – я позволю ему охотиться здесь или бултатыыһыкпын – видимо мне придётся ему разрешить охотиться здесь), нет побудительной формы приобретения, крайне бедна притяжательная форма (например: кинигэлэн – ты имей при себе книгу или саалаахпын – у меня есть ружьё). В русском языке выражение жалости производится исключительно аналитически, а на языке саха – с помощью аффиксов к любым глаголам можно выразить жалость (бараахтаа – ты бедненький иди, хотя тебе тяжело или оонньоохтуом – конечно мне тяжело, но я попытаюсь поиграть), нет побудительной формы совместного действия (барыс – ты иди со мной), мало глаголов, обозначающих многократность выполнения (иһитиннэртэлээ – а ты проинформируй всех по мере возможности).

По сравнению с языком саха в русском языке мало глаголов обозначающих приобретение каких либо свойств (таких, как например: одеревенеть). Кроме того, язык саха чрезвычайно насыщен образными и звукоподражательными словами и к сожалению почти ни один из них не имеет не только перевода, но даже приблизительного описания. Не вдаваясь в дальнейшее перечисление различия языков можно сказать, что в языке саха имя существительное или имя прилагательное могут иметь от трёх до пяти тысяч аффиксов. Чистые глаголы также могут иметь от двух до трёх тысяч аффиксов. В полном разложении современный язык саха представляет около 45’000’000 слов (см. Саха русский компьютерный словарь на 45’000’000 слов [2]). Наиболее существенными недостатками языка саха являются:

- наличие большого количества парных слов;
- наличие большого количества омонимов;
- инфинитив выражается аналитически (методом разрыва логической связи).
- синтаксис якутского языка хотя формально похож на синтаксис других тюркских языков, но тем не менее, изафетные словосочетания в якутских конструкциях отличается от других тюркских языков, тем более с синтаксисом русского языка (имеется существенное расхождение).

Главным отрицательным моментом при переводе является невозможность дословного перевода.

Если на русском языке практически всегда есть или подлежащее, или сказуемое, а также случай, когда оба имеются, то в языке саха из-за особой синтетичности очень сложно найти их (например: ааннаабатаххын – оказывается ты вовсе не установил дверь или дьиэлээбиттээх – в принципе он как-то (один раз) был у себя дома).

Следовательно единственно правильным путём для перевода с языка саха на русский язык является перевод основанный на кодирование смысла предложения в виде математической формулы неизменной (инвариантной) независимо от способа построения и подачи предложения. Но при этом автоматически появляется, как при любом моделировании, вопрос об устойчивости модели при неизвестных словах (аббревиатуры, термины, названия и.т.д.) или омонимах. Особую сложность представляет склонённые или своеобразно составленные фразеологические конструкции.

В данной работе автором будет впервые сделана попытка численно кодировать смысл предложения с точки зрения русского языка на тюркский язык (ни тюркская, ни романская группа языков по однозначности (по лексическому запасу) слов не могут конкурировать с русским языком). Будут созданы специальные динамические библиотеки синтеза русского языка. Тогда для синтеза предложения на русском языке достаточно будет указать номера объектов, их свойства и численные индексы методов. Само

кодирование будет объектным, точно таким - каким является язык саха. Использование реперным языком русского языка позволит дальнейший перевод якутского языка через русский на другие тюркские языки.

Конечно любого гражданина (обывателя) интересует (а часто заранее считают бесполезным и ненужным делом перевод умирающего языка) точность перевода естественного языка. Какие принципы перевода я ставлю перед собой и чем отличается этот принцип от индоевропейских языковых переводчиков. Для тех, которые относятся к скептикам могу напомнить тот факт, когда я стал заниматься орфографией якутского языка (только обиходные якутские слова в разложении принимают 45’000’000 значений, при этом надо было учесть своеобразные склонения неправильных существительных и глаголов) тоже никто не верил в успех, сейчас когда программа по электронной проверке знает около 50 миллионов комбинаций слов с учётом всех нестандартных склонений в теле Word-а [3, 4], могу уверенно заверить, что моя база данных уже содержит определённые семантические коды готовые для дальнейшей аналитической работы. Теперь о точности, я совершенно согласен с Алексеем Сокирко [5], что машинные переводчики нужны для решения каких-то прикладных задач. Давайте подумаем, всегда ли нужно делать точный перевод на английский язык слов хихикнул, фыркнул, усмехнулся, смеялся, засмеялся, рассмеялся, разразился, насмеялся, высмеял, отсмеялся в русском варианте (нагромождать не всегда понятными длинными толкованиями или «выискивать» нанометры при измерении длины ложки). Равно необходимость точного перевода с синтетического языка саха для передачи одной простой мысли на русский язык, например: вода в озере мелеет, при этом надо учесть, что на языке саха, слово озеро в зависимости от формы, размера, глубины и формы берега имеет около 50 названий, иначе ни один древний охотник за тысячи километров не находил бы в тайге положенного места по описанию какого-то старца, ведь не зря в Якутию можно поместить пять Франции, кроме того каждая речушка находясь в невообразимо дальних расстояниях от поселения людей имеет своё старинное название. Ещё лучше, нужно ли иметь сверх хорошую видеокарту для монитора VGA с разрешением 256 цветов. Пока художественный машинный перевод литературного произведения с любого языка на другой язык, глубоко убеждён в этом, не возможен, надеюсь никогда не будет. Каждый язык прекрасен только для своего народа. Художественный перевод должен соответствовать определённым понятиям, этике и обычаям другого народа, в каких-то случаях должно быть умолчание, а других – добавление (разъяснение).

Наиболее близким алгоритмом работы переводчика является синтаксический анализатор описанный в [6], где производится выделение простых предложений в составе сложного и производится построение синтаксического древа, которое потом воспроизводится на синтаксическом древе другого языка. Отличие состоит в понимании понятия смысла предложе-

ния, как логической системы состоящей из трёх объектов, а не в виде синтаксического древа. Заранее оговорюсь, противопоставление алгоритмов для решения языковых структур с различными синтаксическими построениями в корне неверно. Коллеги которые занимаются индоевропейскими языками правы по своему, бог им в помощь.

CS = ObjOfAttention.ObjOfConclusion.ObjOfTime

где ObjOfAttention - Объект внимания,

ObjOfConclusion -Объект-заключение,

ObjOfTime - Объект-время.

Здесь понятие объект условное, грамматически может быть любой частью речи. Кроме того объект может иметь однородное перечисление частей речи. Привязка к грамматике в большей части приводит к тупику (неопределённости), так как во многих якутских предложениях могут отсутствовать, как подлежащие, так и сказуемые. Грамматические правила присутствуют только в методах и свойствах.

Далее Объект может подразделяться на классы и иметь методы и свойства. Синтаксис Объекта внимания –
ObjOfAttention.Class.Method.Property1.Property2

Под классом в данном случае понимается иной смысл слов в отдельности (омонимы), так и в объединении с другими словами (фразеологизмы). Объект и класс состоят из идентификационных порядковых номеров в соответствующих словарях и исходных слов.

Свойствам Объекта внимания могут быть местоимение и имя прилагательное.

Метод есть номер программной оболочки распознавания объекта, класса, трансляции грамматической структуры, а также распознавания образа причины (почему, отчего, по какой причине), а также места (где, куда, откуда). Метод содержит полный смысловой перевод на русский язык Объекта внимания.

Объект внимания всегда есть, даже если переводится одно слово.

Объект-время также содержит класс, методы и свойства, также автономен. Метод объекта времени тоже есть номер программы, которое распознаёт из текста время, а также образы (когда, как долго, с каких пор, до каких пор). Метод содержит на русском языке описание временных факторов. Объект-время не всегда присутствует в описании смысла предложения, если время неопределённое.

Объект-заключение является завершающим объектом смысловой нагрузки, он зависимый от предыдущих объектов и содержит пять основных методов и свойств. Должен содержать в методах перевод на русском языке заключительную стадию образов: зачем, для чего, с какой целью, как, каким образом, в какой степени, подобно как, при каком условии, несмотря на что, по сравнению с чем.

В конечном итоге смысл предложения сводится к числу с фиксированной длиной, составленному аналогично грамматическому построению якутского языка. Такое построение смысла предложения резко облегчает (по скорости) учёт предыдущих предложений для анализа последующих высказываний. Хотя бы для определения рода объекта (он или она), но в основном для исключения омонимов и синхронизации числа объектов. В отличии от русского языка в языке саха нет слов имеющих только множественное значение (очки, брюки и т.д.) или слов имеющих только единственное значение (фанера, вата и т.д.), и нет понятия мужской, женский род.

Важным является создание смысловых библиотек (разрешённых понятий) и библиотек исключения (для дальнейшего поиска в основной базе лексем разрешённых по смыслу). Наличие специальных библиотек по устойчивости модели совместно со смысловыми анализаторами позволяет индицировать на мониторе вероятность точности передачи смысла сказанного.

Не менее важным является синтез предложения на русском языке, особую неприятность для моей работы представляет грамматически правильное синтезирование частиц (например: пришёл из леса или пришёл с горы) для языка саха целеуказание производится синтетически в виде аффиксов и естественно нет равноценных отдельных слов. Конечно, была бы специализированная программа (библиотека dll) русского языка исправляющая грамматические ошибки электронных переводчиков на русский язык со смысловой проверкой конечно было бы идеально. Я не думаю с азиатских языков только якутский язык будет переведён на русский язык. По понятным причинам все стремятся «американизироваться», но самое интересное, пока что разработчики программных средств не поймут обязательного наличия смыслового образцового репера на русском языке, будут продолжаться потешные переводы с одного языка на другой. Продолжающаяся привязка к английскому языку, у которого омонимов больше чем даже у тюркских языков только растягивает время.

По моему мнению, для правильного перевода с любого языка на другой, обязательна цепочка: исходный язык – русский язык – язык на который переводится – русский язык – исходный язык. Понятно, создание смысловой библиотеки русского языка требует огромных усилий и времени (конечно денег), но к сожалению это единственный, надёжный способ правильной передачи мысли с одного языка на другой язык. Любой инженер электронщик подтвердит максимальную устойчивость компенсационной системы перехода с одного уровня на другой уровень (стабилизированные блоки питания построены на этом принципе).

Структура синтетического языка по-своему интересна и для разработчиков машинных языков. В синтетическом языке саха никогда не бывает префиксов, соблюдается строгая последовательность объект – метод –

свойство (агглютинативный принцип). При этом количество методов строго нормировано и все они применимы практически для всех объектов, что очень важно (быалаа – привяжи верёвкой, москвалаа – поезжай в Москву). Количество аффиксальных комбинаций любого метода зависит от модальности (вероятности события) и трёх временных факторов. К свойству относится местоимение или притяжательное местоимение, при этом фонетика языка так построена, что свойство легко стыкуются непосредственно с самим объектом (саам – моё ружьё, саатыттан – от его ружья).

В заключении могу объяснить, почему работа выносится на обсуждение тогда, когда работа ещё не завершена. Данная работа производится исключительно в инициативном порядке при полном безразличии чиновников республики, естественно критические замечания коллег по подобным работам в процессе эволюции только сэкономит время, для немолодого человека этот фактор архиважный.

Литература

1. Е.И. Убрятова «Исследования по синтаксису якутского языка» Изд-во АН СССР, Москва, Ленинград, 1950.
2. «Саха русский компьютерный словарь на 45’000’000 слов с возможностью обратного перевода» Проект №00-06-96209, Арктика 98, РФФИ, под рук. д. фил. н. профессора Е.И. Коркиной
3. В.В. Мигалкин Авторское Свидетельство № 2001611636 «Sakha», Роспатент, 2001.
4. В.В. Мигалкин Авторское Свидетельство № 2001611637 «Sakha Orthography», Роспатент, 2001.
5. А.Сокирко «Будущее машинного перевода» Компьютерра, № 21, 2002.
6. И.Ножов «Синтаксический анализ» Компьютерра, № 21, 2002.