ՀԱՅԿԱԿԱՆ ՍՍՌ ԳԻՏՈՒԹՅՈՒՆՆԵՐԻ ԱԿԱԴԵՄԻԱՅԻ ՏԵՂԵԿԱԳԻՐ
ИЗВЕСТИЯ АКАДЕМИИ НАУК АРМЯНСКОЙ ССР
Հասարակական գիտություններ № 1, 1959 Общественные науки

В. М. Григорян, М. И. Белецкий

Об алгоритме машинного перевода с армянского языка

При Вычислительном центре Академии наук Армянской ССР создана небольшая группа, состоящая из математиков и лингвистов, совместно работающих над составлением алгоритма машинного перевода с армянского языка. Этой группой почти завершено составление аналитической части алгоритма, накоплен опыт (правда, еще очень небольшой) в области формального описания языка, что создало возможность сделать в настоящей статье кое-какие предварительные выводы.

Работа над алгоритмом перевода с армянского языка ведется на основании принципов, разработанных группой А. А. Ляпунова, О. С. Кулагиной и И. А. Мельчука. Естественно, что своеобразие армянского языка (в частности, смешанный морфолого-агглютинирующий характер его системы) привело к необходимости частичных отступлений от имеющейся схемы. В настоящей статье мы попытаемся остановиться на некоторых проблемах, связанных с составлением алгоритма машинного перевода с армянского языка в его аналитической части, одновременно, правда, в самых общих чертах, затронув отдельные вопросы принципа.

Для машинного перевода составляются две группы правил: правила анализа и правила синтеза.

Анализ заключается в обработке морфологических особенностей каждого слова переводимого текста и в установлении тех форм зависимости между каждой парой слов, которые имеют место в данном тексте. При этом этап анализа предполагает перевод реального текста в цифровой код. Это значит, что каждое слово реального текста заменяется информацией к нему, т. е. в распоряжение машины поступает какая-то цепочка цифр, каждая из которых характеризует данное слово с различных точек зрения (часть речи, наличие омонимичного слова, способность образовать производные слова и т. д.).

Помимо этого, первоначальная информация пополняется дополнительными сведениями морфологического и синтаксического характера. Иначе говоря, на основании заданных правил машина сначала анализирует (в определенной последовательности) морфологическую специфику каждого переводимого слова, представленного в виде цепочки цифр, а затем устанавливает синтаксические связи между парами слов, замененных информациями, на основании их взаимной зависимости. В результате этого процесса переводимый текст распадается на ряд двучленов (зависимое и зависящее слово), которые, в свою

очередь, составляют последовательность дискретных эквивалентов данной фразы. Заметим, кстати, что указанные двучлены в существующей терминологии принято называть конфигурациями.

Следующая система правил относится уже к обратному процессу — синтезу. Синтезирующие правила поступают в распоряжение машины с целью перевода цифровой цепочки в реальный текст на переводящем языке. Этап синтеза характеризуется операциями, обратными тем, которые имели место на этапе анализа, так как если при анализе с помощью особо составленного словаря и системы таблиц переводимый текст подвергался формальной записи, то при синтезе осуществляется обратный переход от закодированной записи в запись, свойственную нормализованной графической системе данного переводящего языка.

Машинный перевод предполагает также необходимость промежуточного этапа — этапа перехода, при котором приводятся в соответствие информации переводимого и переводящего языков.

Таким образом, для получения перевода с помощью машины должны быть составлены две группы правил — правила анализа и правила синтеза. Правила анализа предполагают осуществление перевода данного текста в особый код, который был бы „удобен" для машины — в цифровой код. Понятно в связи с этим, что для осуществления такого специфического описания текста необходима методика, коренным образом отличная от принятой в так называемых „традиционных", „описательных" грамматиках. Следовательно, при описании армянского языка с целью сделать его „понятным" машине совершенно неизбежны отступления от общепринятых схем. Так, несколько по-иному классифицируются части речи, вносятся коррективы в ставшие привычными представления о глагольной системе и т. д.

Анализ армянского текста в целях машинного перевода (как, впрочем, и само явление „машинный перевод") не следует понимать слишком прямолинейно и категорично. Наш анализ ведется пока исключительно на материале языка математической литературы. Подробное описание причин, вызвавших предпочтение именно языка математических работ (статей, монографий, учебников и т. д.), не входит в круг задач настоящей статьи и может представить собой тему самостоятельного исследования.

Этап анализа ведется в определенной последовательности. Он начинается с поиска слова в словаре основ.

Основы слов выделяются и записываются в словаре основ, прежде всего, исходя из соображений практического удобства. Для слов, изменяющихся без чередования и выпадания букв, основа выделяется так же, как это делается в традиционной грамматике. Например, у слов *անվանել, կարդալ, թեորեմաներ* основами служат *անվան, կարդ, թեորեմա.*

Для слов, изменение которых сопровождается выпадением гласных, в словарь записываются две или более основ. Так, для односложных

существительных с гласными *ու, ի* выделяются две основы: *թիվ* и *թվ*, *առն* и *ան*. Если происходит чередование конечной согласной основы, то эта согласная вообще не относится к основе. Так, у глаголов *եզրակացնել (եզրակացրած)* и *հասկանալ (հասկացող)* основами будут: *եզրակաց* и *հասկա*.

Большинство префиксов мы также относим к основе. Исключение составляют только те префиксы, которые обладают большей смысловой нагрузкой по сравнению с другими префиксами, что создает возможность для легкого нахождения аналогичной формы в других языках. Таковы, например, префиксы *ան, ամենա*, и т. д.

Как указывалось выше, задачей анализа является замена слова информацией к нему, т. е. цепочкой цифр, полностью характеризующей его значение и форму. Часть этой информации дает основа слова, что создает возможность записи этой части в словаре основ. Информация в словаре основ содержит следующие сведения о слове:

1. Часть речи.

2. Управляющая группа. Содержание этой группы сводится к указанию на то, какими словами и в какой форме может управлять данное слово.

3. Способность слова входить в состав оборота.

4. Наличие в словаре омонимичного слова.

5. Способности слова образовать производные слова.

6. Наличие особенностей, проявляющихся при проведении анализа.

7. Номер лексического значения знаменательного слова или номер служебного слова.

8. Грамматическая форма. Для существительных указывается число и падеж, для глагола—время, число, залог, наличие возвратности и т. д.

9. Наличие отрицания.

Сведения в закодированном виде заносятся в соответствующие графы словаря основ. Содержание каждой из этих граф оказывается существенным на разных уровнях анализа, что будет показано при рассмотрении этих уровней.

Основы в словаре основ расположены в алфавитном порядке, причем основа, включающая другую основу, ставится перед этой последней. Это связано с тем, что слово в тексте проверяется сначала на вхождение в него большей основы, а затем меньшей. В случае нахождения большей основы поиск будет прекращен. Если бы в словаре сначала была помещена меньшая основа, то после ее нахождения следовало бы сделать вторую проверку—не входит ли меньшая основа в большую. Например, в словаре расположены сначала основы *այսինքն* и *այսպես*, а затем *այս*.

Итак, прежде всего каждое слово фразы ищется в словаре основ, т. е. ищется первая из основ словаря, целиком содержащаяся в данном слове, причем в начале его. Такая основа может не быть

найдена за счет того, что слово начинается с префикса. В таком случае слово ищется в словаре префиксов. Найденный префикс отбрасывается и затем слово ищется в словаре основ. Например, слово *կգինի* не будет найдено в словаре основ. Тогда находится префикс *կ* в словаре префиксов, затем основа *գին* в словаре основ, выписывается информация к этой основе и к информации приписываются указания на образование будущего времени.

Таким образом, к концу уровня поиска фраза записана как последовательность информаций к словам и последовательность окончаний, оставшихся после отделения префиксов и основ. Если к основе данного слова имеется омонимичная основа, то при поиске выписывается информация к обеим основам. Например, слова *անվան* и *անվաննել* (соответственно — *имени* — кого-чего и *называть)* имеют омонимичные основы с различными информациями: одна из основ содержит информацию имени существительного, а другая — глагола. Различение омонимии и выбор одной из двух информаций происходит на дальнейших уровнях анализа.

Затем следует *уровень обработки оборотов.*

К оборотам мы относим сочетания двух или более слов, составляющие одну семантическую единицу. Это, разумеется, не может служить точным определением. То или иное конкретное сочетание считается оборотом, исходя из соображений практического удобства при машинном переводе.

Практически оборотами признаются явления типа: *ոչ մի կասկած, ի նկատի ունենալով* и т. п.

Обороты задаются списком — словарем оборотов. У одного из слов оборота (обычно более редко встречающегося вне оборота) в информации дана пометка на способность входить в оборот и указан номер оборота. По номеру в словаре оборотов отыскивается оборот и проверяется—есть ли во фразе другие слова данного оборота. В случае их обнаружения набор информаций к этим словам заменяется информацией, взятой из словаря оборотов. Например, в тексте найдено слово *կասկած*. При нем в словаре основ стоит информация, указывающая на возможность вхождения этого слова в оборот. Одновременно с этим указан и номер оборота. По номеру в словаре оборотов находится оборот *ոչ մի կասկած*. После этого приводится проверка в тексте — стоят ли перед нашим словом слова *ոչ* и *մի*. Если эти слова обнаружены, то информации ко всем трем словам заменяются одной информацией.

К началу уровня морфологического анализа выделена информация, заключающаяся в основе. Задачей этого уровня является выделение и запоминание информации, содержащейся в окончании. Под окончанием понимается все то, что остается от слова после отбрасывания основы.

Окончание может состоять из суффиксов трех родов (как будет видно ниже, к суффиксам мы относим и флексии):

1. Словообразующие суффиксы, присоединяясь к основе, придают слову другое значение и определяют его „часть речи". Сюда относятся, например, суффиксы существительных *ութիւն, իչ, ուհի*, суффиксы прилагательных *ալի, ական* и т. д. Особо учитывается, что у словообразующих суффиксов, как и основ, может быть омонимия, т. е. одинаково записанные суффиксы могут придавать словам различные смысловые значения или определять различные „части речи". Так, суффикс *ական* может явиться суффиксом прилагательного или наречия.

2. К словоизменительным суффиксам мы относим флексии — части слов, указывающие на связь данного слова с другими словами фразы. Таковы падежные и числовые окончания существительных, временные и личные окончания глаголов.

3. Группа основообразующих суффиксов состоит из суффикса возвратности и его вариантов, появляющихся из-за того, что в случае чередования мы относим чередующиеся согласные к суффиксу (см. выше): *նվ, ցվ*.

Перед началом морфологического анализа в информации к основе есть указание на часть речи, которую образует данная основа, если к ней не прибавляется словообразующий суффикс. Если различные словообразующие суффиксы, присоединяясь к одной и той же основе, могут образовывать различные части речи, то в словарь записываются две омонимичные основы, различающиеся информациями.

Если это — глагольная основа, то, прежде всего, ищется основообразующий суффикс. В случае его нахождения в информацию записывается указание на страдательность. Далее (а не для неглагольных основ — с самого начала) производится поиск словообразующего суффикса. Словообразующий суффикс меняет значение многих граф информации, в том числе графы, указывающей часть речи. В информацию заносится также указание на значение, которое придает слову суффикс (оно понадобится для синтеза слова на другом языке). После нахождения одного словообразующего суффикса ищется второй, и так до тех пор, пока не будет найдено все слово в целом.

Покажем на нескольких примерах, как протекает морфологический анализ. Пусть в тексте встретилось слово *գրաղված*. В словаре основ будеть найдена основа *գրաղ* и приписана к ней информация, в которой, в частности, будет указано, что это глагольная основа. Потом в таблице основообразующих суффиксов находится суффикс *վ* и в информацию вписывает указание на возвратность. И, наконец, в таблице словообразующих суффиксов глагола находится суффикс *ած*. После всех проделанных операций наше слово будет считаться отглагольным прилагательным со значением страдательности (по нашему разделению частей речи причастие относится к прилагательным).

Пусть нужно проанализировать слово *գործնականորեն*. В словаре основ находится основа *գործ* и слово считается существительным.

После нахождения суффикса *նական* оно будет оценено как прилагательное и, наконец, после нахождения суффикса *որեն* слово будет отнесено к наречиям. Одновременно с этим происходит изменение и других граф информации.

После поиска словообразующих суффиксов производится поиск словоизменяющих суффиксов. Для существительных ищутся сначала флексии числа *եր* и *ներ*, а затем — флексии падежа. Словоизменяющие суффиксы создают или меняют для существительных информацию о числе и падеже, а для глаголов — о времени, числе, лице и наклонении. Для сложных форм глагола морфологический анализ основного глагола не может дать всю нужную информацию, и она получается на следующем уровне (см. ниже).

Информация к основам и словообразующим суффиксам составляется так, чтобы она указывала форму слова в случае отсутствия словоизменящего суффикса. Так, к слову *թեորեմա* дана информация о том, что это — существительное в именительном падеже единственного числа. Если в тексте встречается слово *թեորեմա*, то оно и остается с такой информацией. Если встречается слово *թեորեմաներ*, то на уровне морфологического анализа меняется информация о его числе. Если же будет встречена форма *թեորեմաների*, то, кроме того, будет изменена и информация о падеже.

Следующая операция проводится на уровне обработки особенностей. При этом проводится обработка слов, имеющих какую-нибудь особенность, отличающую их от подавляющего большинства слов, записанных в словаре основ. Практически такими словами с особенностями являются слова, лишенные самостоятельного значения, которые отличаются служебной ролью и одновременно с этим пишутся отдельно.

В нашем алгоритме анализ форм глагола (в силу их аналитического характера) проводится на уровне обработки особенностей. В этой связи хотелось бы остановиться на рассмотрении глагольной системы армянского языка, которая, в общих чертах, представляет собой следующую картину. Значение времени может создаваться двояким образом: аналитически и синтетически. Причем аналитическое образование связано с употреблением так называемых „дербайных" форм плюс вспомогательный глагол. Наряду с разветвленной и богатой системой аналитически образуемых времен есть одно—прошедшее совершенное (аорист) — время, значение которого передается синтетической формой: к основе (см. выше) прибавляется (в I лице ед. числа) суффикс *ա-ի, ե-ի, ր-ի, ա, ց-ա*. В традиционной грамматике дербайные формы принято отождествлять с причастиями (в том значении, в каком употребляются формы отглагольного прилагательного в русском языке).

Таких дербайных (или по традиции — причастных) форм принято выделять семь:

1. Дербайная форма, способная образовать значение инфинитива (дербай на *ել, ալ — անորոշ*).

2. Дербайная форма, образующая значение имперфекта (дербай на *ում — անկատար*).

3. Дербайная форма, способная образовать значение будущего времени (дербай на *ելու, ալու — ապառնի*).

4. Дербайная форма, способная образовать значение давнопрошедшего времени (дербай на *ել, ցուլ — վաղակատար*).

5. Дербайная форма, способная образовать значение прошедшего результативного времени (дербай на *ած — հարակատար*).

6. Так называемая „подлежащая" форма дербая (дербай на *ող*).

7. Форма дербая, лежащая в основе „отрицательного наклонения".

Дербайные формы 2, 3, 4, 5, которые способны образовать (или образуют) временные значения, входят в систему форм изъявительного наклонения (в которую включается форма аориста). Значение времени, повторяем, образуется прибавлением к дербайных формам настоящего или прошедшего времени вспомогательного глагола *լինել* (быть). Таким образом, изъявительное наклонение в армянском языке представлено девятью „временами". При этом важно отметить, что возникает двуплановость аналитически образуемых времен с точки зрения предшествования (дербай 2, 3, 4, 5 плюс вспомогательный глагол настоящего или прошедшего времени).

Однако в интересах формального описания системы времен изъявительного наклонения оказалось целесообразным рассмотреть дербайные формы в плане возможности их функционирования в сочетании со вспомогательным глаголом или самостоятельно. Дело в том, что наличие вспомогательного глагола при дербайной форме снимает вопрос об омонимии, который в ряде случаев может возникнуть. С точки зрения возможности омонимии дербайные формы делятся на две группы: в одну входят те формы, которые в современном армянском языке невозможны без наличия стоящего справа или слева от них вспомогательного глагола, во вторую группу входят те формы, которые не предполагают обязательного наличия вспомогательного глагола и которые либо способны функционировать в значении определения (т. е. быть собственно причастиями), либо могут явиться омоформами с прямыми или косвенными формами глагольной парадигмы.

Так, мы получаем следующую картину:

Группа 1		*Группа 2*
а) дербай 2 (на *ում*) б) дербай 4 (только с суфф. *ցել*)	дербаи, способные функционировать в значении причастия — дербай 5 (на *ած*)	дербаи, образующие омоформы: а) дербай 3 (на *ելու, ալու*), б) дербай 4 (с суфф. *ել*).

Исходя из всего сказанного, при обработке глагольных форм возможны два случая: во-первых, при обнаружении дербайной формы группы 1 окажется достаточным наличие самого факта окончания на *ում* или *ցել*, во-вторых, при обнаружении дербайной формы группы 2 необходима проверка на наличие вспомогательного глагола.

Следует отметить, что омоформия, характерная для части глаголов группы 2, проливает свет на специфику армянских дербайных форм вообще. Так, одна из форм этой группы — форма дербая на *ած* — в случае употребления со вспомогательным глаголом функционирует во временном (прошедшее результативное) значении; без вспомогательного глагола — это собственно причастие, употребляющееся в функции атрибута. Форма дербая на *լու*, употребляемая без вспомогательного глагола, совпадает с формой родительного падежа склоняемого инфинитива; при этом родительный падеж инфинитива в армянском языке отчетливо напоминает герундив (*կարդալու գիրք* — книга, которая должна и. и может быть прочитана). Дербайная форма давнопрошедшего времени на *ել* (за очень редким исключением) может быть омонимична с формой дербая, способного образовать значение инфинитива. Одновременно с этим формы дербая на *ցել* являются „чистыми" показателями категории времени. Следовательно, само представление о дербайных формах, как о чем-то единообразном и однородном, подвергается уточнению.

Еще более специфично, по сравнению с перечисленными дербайными формами, ведет себя „подлежащный" дербай. Этот дербай принципиально лишен способности образовывать значение времени. Сочетание „подлежащного" дербая со вспомогательным глаголом составит конфигурацию существительного nomina agentis плюс глагол-связка. Однако „подлежащный" дербай может иметь и другое значение — атрибута, выраженного отглагольным прилагательным. „Подлежащный" дербай в сочетании с существительным составит конфигурацию прилагательное-существительное. Говоря о „подлежащном" дербае, таким образом, мы должны констатировать особый вид омонимии с проверкой на синтаксические условия.

Не имея возможности останавливаться на других особенностях, обрабатываемых на этом уровне, перейдем к следующему уровню — синтаксическому анализу.

В основе синтаксического анализа лежит разбивка фразы на типовые конфигурации двух информаций. Конфигурацией считается всякая пара слов, связанная по смыслу. Понятно, что этот критерий не совсем удачно сформулирован, но практически он очень удобен. Конфигурации задаются тремя параметрами:

1. Из чего составлена конфигурация, из какой пары информаций.

2. Каково взаимное расположение этих информаций.

3. Какими классами слов разделяются информации данной конфигурации.

Тип конфигурации выделяется на основании совпадения всех трех параметров. Практически оказывается достаточным совпадение первых двух параметров. Так, одной конфигурацией окажутся *կարմիր մատիտ* и *կարդալու գիրք*. Эти типовые конфигурации поступают в распоряжение машины списком, в котором они нумеруются; другими словами, нумеруются строки в особой таблице конфигураций. Конфигурации выделяются на основании отношения управления. С целью выяснения групп управления, нами была проделана следующая работа. Во-первых, был составлен список глаголов, сильно управляющих падежом без послелога (например, глагольная группа 1 — глаголы, сильно управляющие винительным беспослеложным). Затем были рассмотрены глагольные группы, сильно управляющие падежом с послелогом или предлогом (например, глагольная группа 14 — глаголы, сильно управляющие родительным падежом с предлогом *ըստ*) и т. д. После разбивки слов по типу управления нами выделены определенные управляющие группы. Надо сказать, что при этом также были обнаружены отдельные черты, характеризующие синтаксические отношения в армянском языке. Например, было установлено, что все глаголы принципиально способны управлять местным падежом; оказалось, что глаголы движения, которые в других языках невозможны в качестве управляющих винительным беспредложным, в армянском языке могут требовать формы именительного-винительного падежа беспослесложного. Причем в этой форме при глаголах движения возможно только существительное (а не местоимение).

Выше было дано описание алгоритма анализа армянского текста. Теперь в самых общих чертах остановимся на способах записи этого алгоритма и реализации его в машине. Разумеется, здесь будет идти речь о таких средствах формального описания и машинной реализации, которые оправдывали бы себя не только при анализе армянского текста, но и на всех этапах перевода с любого языка на любой.

Как уже указывалось, в каждый момент перевода мы имеем дело с кусками слов и поставленными в соответствии им упорядоченными наборами цифр-информациями. Оставшийся к данному моменту кусок слова назовем *слогом,* а совокупность слова и информации *слог-информацией. Фраза* представляет собой упорядоченную последовательность слог-информаций. Таким образом, машинный перевод сводится к процессу преобразования фраз.

В связи с этим возникает задача разработки математического аппарата, описывающего действие над фразами. Он будет в какой-то степени аналогичным нормальным алгоритмам А. А. Маркова[1].

Над фразами задается некоторый набор элементарных операций двух типов:

---

[1] А. А. Марков, Теория алгоритмов, Труды Математического института АН СССР, т. XLII.

1. Операции изменения слог-информации или фразы в целом (прибавление или вычитание слога, изменение информации, устранение или введение новой слог-информации, изменение порядка слог-информаций во фразе и т. д.).

2. Операции проверки (проверяется, стоит ли определенное число в определенной графе информации, есть ли данный слог в таблице, стоит ли рядом с данной информацией, удовлетворяющей определенным условиям, и т. д.). Операции проверки сами не изменяют фразу, но указывают, к какой из операций изменения переходить.

Таким образом, алгоритм любого этапа перевода с любого языка на любой (а также отдельных уровней этого этапа) представляет собой последовательность операций над фразами. На этапе анализа любая фраза языка переводится в последовательность информаций, на этапе синтеза — осмысленная последовательность информаций во фразу языка.

Кроме того, каждый этап перевода с конкретного языка жестко связан с определенным набором таблиц. Для этапа армянского анализа это были: словарь армянских основ, словарь оборотов, таблицы префиксов, суффиксов и т. д. Для удобства записи операций эти таблицы нумеруются набором индексов, каждый из которых характеризует содержание или назначение таблицы. Фраза, как набор слог-информаций, может быть задана электронно-вычислительной машине. Буквы слога могут быть легко закодированы двоичным кодом, а перевод информации в двоичную систему также не представляет никакого труда. Операции над фразами могут быть заданы как последовательности элементарных операций машины. Таким образом и осуществляется перевод на существующих сейчас вычислительных машинах. Но ввиду того, что эти машины специально не предназначены для перевода, операции над фразами реализуются на них слишком сложно, память не позволяет включать широкий набор таблиц. Поэтому машинный перевод на этой стадии имеет только экспериментальное значение.

Из сказанного выше можно легко заключить, что в работе над машинным переводом необходима слаженная и хорошо координированная работа математиков и лингвистов. Машинный перевод ставит много новых и интересных проблем, решение которых носит специфический характер. Было бы очень хорошо, если бы указанное сотрудничество проводилось не только в рамках коллектива людей, непосредственно работающих над составлением алгоритма.