COMPUTATIONAL LINGUISTICS

V

COMPUTING CENTRE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST, 1966

COMPUTATIONAL LINGUISTICS
V.

COMPUTING CENTRE OF THE HUNGARIAN ACADEMY OF SCIENCES
BUDAPEST, 1966

Publisher: The Computing Centre of the Hungarian Academy of Sciences
Address: Budapest I., Uri utca 49

Felelős kiadó: FREY TAMÁS
Készült a Muzeumok Rotaüzemében, 25 iv terjedelemben,
420 példányban. Szám: KK-203/1966.

C O N T E N T S

M I S C E L L A N E A

R E V I E W S

РАБОТЫ ЛМП МГПИИЯ ПО СЕМАНТИКЕ

Ю.Д. Апресян, К.И. Бабицкий

В последние десятилетия в лингвистике разрабатывались модели двух типов: модели речевой деятельности человека и модели исследовательской деятельности лингвиста. Модель первого рода представляет собой либо процедура /алгоритм/, имитирующую способность человека воспринимать и производить текст, либо исчисление, формализующее его интуитивное представление о правильном и неправильном в языке. Модель второго рода - это процедурк /алгоритм/, имитирующая процесс, в результате которого лингвист от наблюдаемого сырого материала приходит к представлению о способе его организации.

Назначение моделей второго рода состоит в том, чтобы внести элемент объективности в выбор тех содержательных представлений, из которых исходит исследователь при создании модели первого рода. Вообще говоря, модели первого рода не нуждаются в таком обеспечении. Другим и, может быть, более надежным способом обнаружить правильность модели первого рода является экспериментальная проверка ее эффективности. Поэтому в лингвистике модели первого рода часто строятся независимо от моделей второго рода.

До недавнего времени среди разрабатывавшихся моделей первого рода абсолютно преобладали синтаксические и морфологические. Семантика языка была представлена в этих работах исключительно в виде системы грамматических значений. Практика работ по машинному переводу вскоре убедила исследователей в невозможности эффективно описать язык без детальной разработки его семантической системы, и были предпрняты попытки построить семантическую модель первого рода.

Решающий шаг на этом пути сделан в работах ЛМП [1], авторы которых первыми ясно сформулировали и частично реализовали мысль о том, что адэкватная модель первого рода должна имитировать владение значением слов.

В блестящем предисловии к сборнику, написанном А.К. Жолковсим, выдвинут тезис о том, что владение смыслом слов проявляется у говорящего в способности по-разному выразить одну и ту же мысль, а у слушающего - в способности понять семантическое тождество внешне различных высказываний.

Ясное понимание этого основного принципа и вытекающих из него задач и определяет замысел и программу работы авторов сборника.

В прадлагаемом разборе сборника взгляды авторов изложены нами в том виде, как нам удалось из понять, при том по возможности связно и компактно. Поэтому в нашем изложении не следует искать текстуальной близости /за исключением особо оговариваемых случаев/, хотя мы надеемся, что смысл работ сохранен.

Когда говорят о разных способах выражения одной и той же мысли, или /что то же самое/ о семантическом тождестве внешне различных высказываний, имеют в виду, что существует некий не данный в прямом наблюдении "язык мысли", или "семантический язык". Если допустить существование такого языка, то производство осмысленного предложения можно представить как перевод с семантического языка на естественный, а понимание предложения - как перевод с естественного языка на семантический. Очевидная возможность по-разному выразить одну и ту же мысль означает, что у некоторого выражения семантического языка есть несколько переводов на естественный язык.

На семантическом языке мысль имеет единственный стандартный способ записи, и если на естественном языке она выражается несколькими различными

способами, которые мы признаем равнозначными /равносильными данной записи/, это в общем случае значит, что на разные слова приходятся разные части выражаемой мысли. Фразы Он недомогает и Он плохо себя чувствует [2] выражают одну и ту же мысль, причем та ее часть, которая в первой фразе выражается одним словом, распределена во второй фразе между тремя различными словами. Из этого следует, что значение слова не является в общем случае элементарной семантической единицей. Оно делимо на элементарные смысли, которые, по предположению, и являются словарными единицами семантического языка. Таким образом, существование более чем одного перевода семантического выражения на естественный язык объясняется предположением, что семантическое выражение построено из единиц более простых, чем значения слов естественного языка; та или иная комбинация таких единиц дает то или иное слово естественного языка. Ясно, что небольшое число таких единиц /"элементарных смыслов"/ дает очень большое число возможных комбинаций, реализуемых словами естественного языка.

Попробуем теперь представить себе, каким способом выражение семантического языка построено из элементарных смыслов. Рассмотрим с этой целью предложения Это заставляет меня уйти и Я вынужден уйти из-за этого. Очевидно, что они равнозначны; факт их равнозначности подтверждается, в частности, тем, что им обоим может быть сопоставлено выражение Это - причина того, что я ухожу, где та же мысль выражена в более явной форме. В составе рассматриваемых выражений имеются некоторые общие части /я, это, уходить/ и части, которыми они отличаются друг от друга /заставляет в противоположность вынужден/. Кроме того, во втором предложении по сравнению с первым существенным образом изменяется порядок следования частей. Это изменение нельзя считать зависящим от слов, общих для двух предложений /я, это, уходить/; остается предположить, что оно связано с заменой слова заставляет словом вынужден. Таким образом, слова заставляет и вынужден ведут себя в некотором смысле как господствующие: от их выбора зависит порядок остальных слов, которые оказываются в положении подчиненных. Заметим, что слово заставляет является господствующим для частей это и меня уйти. Нетрудно сообразить, что слово уйти выполняет ту же роль по отношению к слову я.

Рассматривая наши предложения как равнозначные, мы открываем в строении предложений естественного языка особенности, не совпадающие с привычными грамматическими признаками /часть речи, тип синтаксической связи и т.п./. Это: 1/ деление частей предложения на господствующие и подчиненные и 2/ порядок подчиненных частей. Ясно, что этими признаками и должна характеризоваться структура семантического выражения. Одним из простых способов изобразить выражение, для которого существенны эти признаки, является скобочная запись. Так, выражение, ближе всего подходящее к семантической записи наших предложений, будет иметь вид:

причина (это, уходить (я)).

Таким образом, семантический язык имеет свои слова /элементарные смыслы/ и свой синтаксис /скобочную запись/. Чтобы понять предложение, т.е. перевести его на семантический язык, нужно иметь естественно- /скажем, русско-/ -семантический словарь и правила перевода. К составлению этих двух документов и сводится в первую очередь задача создания модели речевого поведения человека, владеющего значением слов.

Авторы сборника полагают, что для решения указанной задачи необходим еще один документ - список "законов действительности", знанием которых человек бессознательно пользуется при чтении текста и обращении к словарю" [3].

С нашей точки зрения, однако, эта третья задача является вспомогательной, что мы постараемся показать ниже.

Мы начнем рассмотрение сборника с анализа "словарных" работ, а затем разберем работы, касающиеся правил перевода; в заключение будет дана общая оценка всей книги.

Отдельное слово получает свое истолкование лишь в результате рассмотрения большой совокупности слов, относящихся к единой "типовой ситуации". Это - область действительности, в которой действуют переменные объекты, находящиеся в некоторых отношениях друг к другу, причем отношения связываются определенными "законами действительности", по терминологии авторов. Предметом описания являются слова, обозначающие отношения. Значением каждого такого слова считается некоторая ситуация, являющаяся частью типовой ситуации. Все слова, являющиеся ее именами, получают одно общее определение через элементарные смыслы.

С алгебраической точки зрения определение слова представляет собой /по крайней мере, в идеале/ форму особого рода. Множество значений такой формы - это множество конкретных ситуаций. Поскольку семантическая запись предложения есть /сложное/ имя конкретной ситуации, определение слова /или его семантическая запись/ в общем случае отличается от нее только тем, что на месте имен конкретных предметов здесь стоят переменные [4].

Определение для того или иного слова подбирается в ходе содержательных рассуждений, которые авторы называют "портретированием действительности". "Портретирование" состоит в последовательном описании "картинок", которые в совокупности составляют типовую ситуацию. Иллюстрацией может служить следующий пример А.К. Жолковского. Цель - это положение вещей, являющееся желательным для некоторого лица А. Цель отличается от мечты наличием реальных путей к ее осуществлению, а от прочих желаний - тем, что лицо А само ее осуществляет, используя при этом имеющиеся в его распоряжении ресурсы. Для достижения цели лицо А действует или планирует действия, которые кажутся ему целесообразными. Аналогичными рассуждениями отыскиваются и характерные для данной ситуации законы действительности, например, следующие два принципа целесообразной деятельности: 1/ из ряда несовместимых событий лицо А добивается осуществления того, которое более желательно для А, чем другие; 2/ при всяком множестве событий лицу А желательно располагать большим количеством ресурсов [5]. Знание законов действительности необходимо для отыскания правильного определения значения. Судя по тексту сборника, никакой другой цели, кроме этой чисто эвристической, оно не служит, и поэтому "законы действительности" нельзя признать существенным компонентом модели.

Статьи сборника, относящиеся к "словарной" части, строятся по следующему плану. Часть статьи отражает процесс "портретирования". Здесь содержательным образом описывается некоторая область действительности, указываются основные отношения в ней и связывающие их законы. Далее следуют результаты словарной работы, т.е. своего рода материалы к русско-семантическому словарю. Они представлены в виде трех списков: 1/ списка элементарных смыслов, 2/ списка промежуточных понятий с определениями через элементарные смысли, 3/ списка русских слов данной группы с их определениями через элементарные смыслы, промежуточные понятия, а также любые уже определенные слова. Входом словарной статьи является группа слов - имен одной и той же ситуации. Эти слова могут принадлежать к различным частям речи; существенно то, что они находятся либо в отношении синонимии, либо в отношении конверсии [6]. Сло-

варные статьи объединяются в разделы, так что внутри одного раздела входы всех статей содержат нечто общее в значении. Это позволяет провести нетривиальное сопоставление сходных по значению слов и выяснить неожиданные семантические связи.

К числу работ, представляющих собой материалы к русско-семантическому словарю, относятся статьи Н.Н. Леонтьевой /3/, Ю.К. Щеглова /4/, А.К. Жолковского /5/, В.Ю. Розенцвейга /6/, а также статьи Ю.А. Мушанова /9/, Н. Н. Леонтьевой и С.Е. Никитиной /10/, которые по материалу несколько отличаются от названных выше работ.

Из этих работ наибольший интерес представляют статьи Ю.К. Щеглова и А.К. Жолковского, связанные друг с другом и тематически.

В статье Ю.К. Щеглова описываются две группы слов русского языка - "лексика силы" и "лексика, связанная с понятием воли". Описанию значений слов предшествует модель /"портрет"/ соответствующего куска действительности, которая выглядит приблизительно следующим образом. Сила рассматривается как некая субстанция, которая обладает, в частности, свойствами иметь объем, находится в одном из двух состояний - активном или пассивном - и принадлежать к одному из нескольких качественно различных видов. Каждый вид силы /физическая, мыслительная, зрительная и т.п./ имеет свой орган, через который она реализуется в форме деятельности. Специфическая деятельность данного органа предполагает наличие у него запаса пассивной силы, запаса активной силы [7], средств для перевода силы из пассивного состояния в активное и устройств, осуществляющих самую деятельность, т.е. освобождающих, расходующих активную силу. Интенсивность расходования активной силы тем выше, чем больше ее запас. При этом существует некая норма запаса активной силы и, следовательно, интенсивности деятельности.

Деятельность органа - это, с другой стороны, воздействие освобождаемой им силы на некоторый объект. Когда на один и тот же объект действуют две силы,направленные по одной прямой,то они либо складываются,если они направлены в одну сторону, либо вычитаются, если они направлены противоположно, давая результирующую силу, которая и определяет, что произойдет с объектом.

Силы отличаются не только качеством, но и порядком. Сила высшего порядка воздействует на устройство, расходующее активную силу низшего порядка. К числу сил высшего порядка принадлежат разум, воля и чувство.

Разум оценивает все возможные способы поведения с точки зрения целесообразности, составляет множество программ поведения, целесообразность которых не ниже нормы, и которые могут отличаться друг от друга по степени целесообразности и трудности, и для каждой программы указывает величину силы, способной привести ее в действие. Если в данной ситуации не существует ни одного возможного способа поведения, целесообразность которого не ниже нормы, то составляется программа для того способа, целесообразность которого хотя и ниже нормы, но максимально к ней приближается /это дает возможность истолковать значение выражения вынужденные действия/.

Чувство, которое в зародыше имеет функции и разума и воли, оценивает все возможные способы поведения с точки зрения их приятности, составляет множество программ поведения, отличающихся друг от друга по степени приятности и трудности, и для каждой программы указывает величину силы, способной привести ее в действие. Затем воля и волевой компонент чувства вступают во взаимодействие, причем чувство может мешать воле /тогда происходит вычитание сил/ или помогать ей /тогда происходит сложение сил/. Качество результи-

рующей силы определяет, будет ли программа выбираться из "разумного" или из "эмоционального" множества программ, а величина результирующей силы определяет выбор конкретной программы: выбирается самая целесообразная /или приятная/ из посильных программ.

Нарисованная Ю.К. Щегловым картина отрезка действительности, описываемого данной группой слов, в своих основных чертах действительно выражает ту "наивную физику", которая лежит в основе системы значений слов этой группы. Однако, существуют некоторые интересные слова, например, слово компромисс, значение которых требует для своего объяснения определенного усложнения этой картины. Предложенная Ю.К. Щегловым "наивная физика" предполагает, в частности, что принимаемая к исполнению программа является либо чисто разумной, либо чисто эмоциональной. Это представление объясняет лишь один частный случай употребления слова компромисс /бездействие, наступающее при равенстве противоположно направленных воли и чувства/; не объясняется тот более обычный случай, когда в компромиссной программе имеются как эмоциональные, так и разумные составляющие. Для более полного истолкования значения этого слова необходимо ввести в "наивную физику" представление о сложении сил по правилу параллелограма.

Некоторые другие слова, имеющиеся в словаре Ю.К. Щеглова, объяснены, по нашему мнению, неверно, а их правильное толкование также требует расширения рамок предложенной им "наивной физики". Последняя предполагает наличие в аппарате силы "хранилищ" для пассивной и активной сил и "насосов" для "перекачки" силы из первого во второе и для освобождения активной силы. Между тем, для объяснения такого слова, как перенапряжение, требуется добавить к этому аппарату еще некоторый источник всякой силы /т.е. обмен веществ в организме/, а также "насос", пополняющий запас пассивной силы из этого источника. С другой стороны, предполагается, что сила высшего порядка воздействует лишь на один орган аппарата силы низшего порядка - на "насос", освобождающий активную силу. Но поскольку ясно, что при перенапряжении пополнение пассивной силы из источника форсируется волей, следует признать воздействие сил высшего порядка на соответствующий "насос".

Недостатками "наивной физики" объясняется и неудачное с нашей точки зрения толкование слов капризность, легкомыслие и непостоянство, которые в работе Ю.К. Щеглова считаются синонимами и получают одно и то же определение: "Частая смена решений, еще до их выполнения /вследствие смены мнений и желаний/, причем без воздействия извне /влияния, уговоров, принуждения/". По нашему мнению, легкомыслие и капризность - это свойство принимать решения без учета объективно решающих условий обстановки, причем в первом случае выполняется программа разума, а во втором - чувства. В значение слова непостоянство входит еще фактор времени. Если по прошествии некоторого отрезка времени изменились несущественные условия обстановки и решение изменено, то это происходит вследствие легкомыслия или капризности и является проявлением непостоянства. Заметим, кстати, что, наоборот, неизменность решения при существенном изменении обстановки есть проявление упрямства /такое понимание последнего слова также кажется нам более точным, чем предлагаемое автором/. Таким образом, "наивная физика" должна быть обогащена фактором времени, позволяющим обнаружить новые связи в значениях слов.

Сказанным мы не имеем в виду умалить неоспоримые достоинства нарисованной автором картины. Его работа остается наиболее интересной попыткой связать значения обширной группы слов в стройную систему, основанную на дейст-

вительно спрятанной в языке "физике".

Специфика "словарных" работ не позволяет нам сколько-нибудь подробно рассмотреть приводимый в статье Ю.К. Щеглова материал. Нельзя, однако, не отметить некоторых чрезвычайно удачных объяснений, к числу которых относится толкование значений слов давление, сдерживать, сопротивляться. Оказывать давление на лицо А значит действовать "с целью вызвать трудность для лица А несовершения какого-либо действия или неперехода в какое-либо состояние". Значения двух последних слов различаются весьма тонким образом, но построенная Ю.К. Щегловым система оказывается достаточно сильной для того, чтобы "взять" это различие. Сдерживать действия лица В значит "действовать с целью вызвать для В трудность в каузации того, что является целью действий В"; сопротивляться значит то же, что и сдерживать, но при этом действие А включает напряжение.

Многие из понятий, используемых в статье Ю.К. Щеглова, определяются в работе А.К. Жолковского /5/, которая, впрочем, поставляет определения исходных понятий и для многих других статей сборника. В статье А.К. Жолковского описывается лексика целесообразной деятельности - около 150 слов русского языка со значением причины, цели, средства, помощи, готовности и т.п.

В отличие от статьи Ю.К. Щеглова, работа А.К. Жолковского не содержит развернутого изложения "наивной физики"; в ней лишь кратко упомянуты 2 "закона" той типовой ситуации, которая описывается интересующими автора словами; это - принцип целесообразной деятельности /из ряда несовместимых событий А добивается осуществления того, при котором А располагает большим количеством ресурсов/ и свойство силы /действие силы эквивалентно изменению/. Очевидно, при работе над толкованием слов А.К. Жолковский имел в виду совершенно определенную модель действительности, материалы которой были использованы в тексте словарных статей. При желании ее можно было бы реконструировать по этим текстам, но в этом нет особой нужды, так как роль "наивной физики", как мы помним, чисто эвристическая.

В качестве исходных /неопределяемых/ выражений автор использует слова множество, предмет, свойство, отношение, время, пространство, не, и, или, необходимо, достаточно, истинно и т.п., всего 23 слова; с их помощью определяется около 50 промежуточных выражений, в число которых входят и некоторые "идеальные" выражения, не являющиеся словами русского языка /никакой, другой, только, несовместимо, предшествует, соприкасаться, связывать по чему-либо, каузировать и т.п./.

К числу важнейших промежуточных понятий относятся понятия путь, нужно, каузация. Путем к Р называется множество фактов Х, достаточное для Р и такое, что каждая его часть необходима, чтобы Х было достаточно для Р. Заметим, что это определение плохо сформулировано; видимо, фактически имелось в виду следующее: Х - путь к Р, если, и только если, Х достаточно для Р, и ни одна часть Х не достаточна для Р. Нужным для Р называется все то, что является частью пути к Р. Говорится, что Х непосредственно каузирует Р, если, и только если, 1/ Х необходимо и достаточно для Р, 2/ Х соприкасается с Р, 3/ соприкосновение Х с Р необходимо для истинности 1/ /заметим, что этот пункт полностью избыточен/, 4/ Х имеет место не позже Р. Х каузирует Р значит, по определению, что имеет место множество фактов М, и Х, и Р, причем Х необходимо и достаточно для Р в М /этот пункт также избыточен/, и от Х к Р идет цепочка непосредственных каузаций. Каузация отличается от зависимости тем, что М, Х и Р имеют место. Следовательно, каузация есть частный случай зависимости, а именно - реализованная зависимость.

Слова, толкование которых составляет содержание статьи, распределены по четырем рубрикам: 1/ цель, план, роль, итог; 2/ помощь, использование, ресурсы; 3/ сила, давление, преодоление; 4/ рассчитывать, надеяться, готовить, ждать. В лучших определениях отражен не только глубокий анализ значения данного слова, но и тонкое понимание его семантических связей с другими словами. Примеры: 1/ А пренебрегает В ≡ А исходит из того, что роль В в Р очень мала; А игнорирует В ≡ желательность для А считать роль В малой каузирует А пренебрегать В; 2/ А старается сделать Р ≡ А по плану С прилагает много или все ресурсы, нужные для Р; А пытается сделать Р ≡ А старается сделать Р, не имея истинного мнения о том, является ли его план С путем к Р; А пробует сделать Р ≡ А пытается сделать Р; А считает, что его попытка каузирует либо Р, либо истинное мнение, что Р не годен. Образцово описаны рассчитывать в отличие от полагаться, сотрудничество в отличие от взаимопомощи /сотрудничество есть такой частный вид взаимопомощи, когда цели полностью совпадают/, вышло в отличие от случилось, удача в отличие от достижения, успеха и везения, противодействие в отличие от сопротивления, готовность в отличие от ожидания /ждать значит сохранять готовность/ и многие другие.

Менее удачно определены слова исчерпать и поглотить, жалеть и жалко, и некоторые другие. Исчерпать и поглотить считаются синонимами; между тем, не существует, видимо, контекстов, где бы они могли заменять друг друга. Отметим, в частности, что исчерпывает ресурсы лицо, а поглощает ресурсы вещь. Вопреки мнению А.К. Жолковского, слово жалеть в контексте лицо А жалеет В не синонимично словам жалеть и жалко в контекстах лицо А жалеет, что В и лицу А жалко, что В.

Статьи Н.Н. Леонтьевой о словах со значением времени /3/ и В.Ю. Розенцвейга о лексике имущественных отношений /6/ построены аналогичным образом. Однако, предложенные ими "наивные физики" содержат существенные недостатки.

"Наивная физика" времени, предлагаемая Н.Н. Леонтьевой, нарочито усложнена в результате отказа от использования понятий момента /точки времени/ и скорости. Возможно, это вызвано желанием не выйти за пределы "наивных" физических представлений. Нам кажется, однако, что представления о моменте и скорости совершенно не чужды "наивной физике", как она обнаруживается в значениях слов. Это видно из того, что некоторые громоздкие, неестественные и подчас ошибочные определения, содержащиеся в работе Н.Н. Леонтьевой, приобретают простоту, ясность и убедительность, как только мы воспользуемся этими понятиями /ср. толкования слов медленно, быстро, начало, конец, момент, мгновение и др./. С другой стороны, как это ни странно, автор ссылается на "вероятности". Однако, строгие представления, связанные с этим далеко не "наивным" понятием, фактически не используются, так как для толкования соответствующих слов вполне достаточно используемого в сборнике понятия "нормы".

Н.Н. Леонтьевой можно также поставить в упрек отсутствие в ее работе таких очевидно "временных" слов, как заблаговременно, своевременно, безвременно, разбор которых помог бы толкованию слов предварительно, преждевременно, досрочно. Последовательные Т не значит, в общем случае, "отрезки времени, у которых начало одного совпадает с концом другого" /ср. последовательные интервалы/. Фактически определено не значение слова последовательный, а значение "идеального" слова, отсутствующего в русском языке, но необходимого для определения значений многих слов.

Предложенная в статье В.Ю. Розенцвейга "наивная физика" имущественных отношений не свободна от некоторых противоречий. В первом пункте этой статьи

содержится намек на то, что в имущественные отношения входят пары вида человек - ценность; однако, в примечании к этому пункту говорится, что имущественные отношения имеют место между людьми; наконец, в пункте 2.1. содержится замечание о том, что в имущественные отношения входят четверки элементов вида человек - человек - вещь - вещь /то есть отмечен еще один частный случай/. Следовало, повидимому, сказать, что в общем случае имущественные отношения - это n-местные отношения на множестве людей и ценностей /объек - тов/; впрочем, как справедливо замечает В.Ю. Розенцвейг, все они могут быть описаны через одно двухместное отношение, в которое входят пары вида человек - ценность. В примечании 2 говорится о свойстве отчуждаемости - неотчуждаемости ценностей. Фактически то же самое свойство фигурирует в пункте 3.3. под новым именем атрибутивности /неотъемлемости/.

В статье отсутствуют такие очевидные "имущественные" слова, как терять, утрачивать, сохранять; копить, экономить, наживать, тратить, изымать, платить, рассчитываться, завещать; доход, скупой, жадный, щедрый, мот, выгода, хозяин; собственность, имущество. Отсутствие этих слов мешает реализации одного из главных преимуществ семантического словаря - возможности различить близкие синонимы.

Кроме того, в статье нет многих более сложных "имущественных" слов, таких как кредитор, кредит, заработок, нанимать, арендовать, арендатор, наследование и т.п. Для правильного описания подобных слов потребовалось бы расширить рамки "наивной физики" введением таких понятий, как право, использование, различением юридических и физических лиц, реальных и идеальных ценностей [8] и пр. Вообще нам представляется, что система понятий, связанных с имущественными отношениями, в силу известной исторической практики разработана в языке чрезвычайно подробно и тонко, и поэтому "наивность" этой системы не следует преувеличивать.

Некоторые определения не соответствуют тому множеству ситуаций, в которых реально употребляется данное слово. Очевидно, например, что слово лишать не является частным случаем отнимать, так как в значение первого слова не входит признак рефлексивности, являющийся существенным элементом второго /отнимать значит брать в свою пользу; кстати, для слова отнимать этот признак не указан/. Слово дарить значит не только "каузировать безвозмездно обладание некоторой ценностью для другого лица" но и "каузировать неимение этой ценности для деятеля".

Вместе с тем статьи Н.Н. Леонтьевой и В.Ю. Розенцвейга содержат интересные результаты. В частности, в статье Н.Н. Леонтьевой устанавливается тонкое отличие непрерывно от всегда, иногда от редко, еще от уже, досрочно от преждевременно, потом от впоследствии, когда от одновременно и т.д. Выясняются неочевидные семантические связи между словами; в этом отношении особенно интересна пара слов временно и пока: временно А значит, что А относится к некоторому отрезку времени, такому, что следующие за ним отрезки нормально не включают А; пока А, В значит, что А временно и А и В одновременно. В статье В.Ю. Розенцвейга удачно определены слова воровать, одалживать, занимать, возвращать, возмещать, компенсировать и ряд других. Ср., например, следующее определение: В одалживает, А занимает; А возвращает ≡ Ценность принадлежит В; В дает ценность А в момент времени T_0; ценность - ресурс, не отчуждается; А берет ценность у В в T_0; А дает ценность В в T_1.

К работам, представляющим собой материалы к русско-семантическому словарю, примыкают, как было сказано, еще две статьи: Н.Н. Леонтьевой и С.Е.

Никитиной /10/ и Ю.А. Мушанова /9/.

В статье Ю.А. Мушанова описываются виды логических и иных связей, именами которых являются союзы и частицы *или*, *же*, *а*, *но*, *даже*, *только*, *и*, *еще*, *уже*. Значение союзов *или*, *и* сводимо, по мнению автора, к значениям чистых логических связок: *или* обозначает дизъюнкцию, контраваленцию /исключительную дизъюнкцию/ и штрих Шеффера /последнее положение, весьма нетривиальное, к сожалению, не подтверждено примерами/. Значение *и* - конъюнкция. В значениях других слов /*же*, *а*, *но* и пр./ помимо логических компонентов содержатся модальные компоненты, главным образом, значение ожидания. Частица *же* значит, что 1/ знание субъекта не совпадает со знанием наблюдателя, 2/ наблюдатель сообщает об этом субъекту, 3/ последний воспринимает эту информацию как неожиданность. Тот же модальный компонент - неожиданность - обнаруживается в значении союза *но*, который с логической точки зрения является конъюнктивным Союз *а* ближе к союзу *и*, чем *но* /остается, правда, неясным, в каком отношении/. Слово *даже* значит " больше ожидаемого", а слово *только* - "меньше ожидаемого". *Еще* и *уже* употребляются в тех случаях, когда описываемое событие состоит в последовательном появлении двух фактов, причем появление какого-то из этих фактов в данный момент не ожидается. Если не ожидается появление второго факта, употребляется *уже*, в противном случае - *еще*. Для предложения *уже нет* событие состоит из фактов *есть*, *нет*, а для предложения *еще нет* - из фактов *нет есть*. Этот анализ объясняет, почему невозможны фразы *еще поздно* , *уже рано*.

Следует с сожалением отметить, что интересные и содержательные результаты Ю.А. Мушанова в известной степени обесцениваются неряшливым формальным аппаратом, который в работе такого типа требует к себе особенно серьезного отношения.

Статья Н.Н. Леонтьевой и С.Е. Никитиной посвящена описанию значений другого типа служебных слов, а именно - предлогов. Способ определения и классификации значений предлогов, позволяющий осуществлять анализ предложных конструкций при семантическом машинном переводе, излагается на примере предлога *за*.

В разных конструкциях предлог *за* имеет разные значения. Считается, что предлог имеет значение в данной конструкции, если последнюю можно заменить синонимичным ей выражением, содержащим все ее лексемы и не содержащим этого предлога, например, *пройти за день 20 км* - *пройти в течение дня 20 км*. Если такая перифраза невозможна, предлог не имеет значения, ср. *заступиться за товарища* - *встать на защиту товарища*.

Значение предлога, реализуемое в данной конструкции, связано с заменяющей последнюю перифразой. Авторы устанавливают девять стандартных перифраз, выделяя таким образом девять типов значения предлога *за* [9], например, *указание на расположение точек* /*жить за версту отсюда*/, *указание на причину* /*похвалить за смелость*/, *указание на обмен* /*купить книгу за два рубля*/ и пр.

Для того, чтобы при анализе решить, в каком из значений употреблен предлог, служит приводимый в работе алгоритм, использующий семантическую информацию, которая приписана лексемам, входящим в предложную конструкцию. Эффективность алгоритма ограничена многими условиями, в частности, неразрешимостью омонимии эллиптических конструкций. Приводятся интересные примеры такой омонимии однако, им не дается сколько-нибудь последовательного истолкования. Правда, это место не составляет существенной части данной работы, которая выполнена, в основном, весьма добросовестно и по-деловому.

Заслуживает одобрения, что авторы снабдили текст сборника, так сказать, "конкордансом" в виде алфавитного указателя всех слов, так или иначе толкуемых в работах сборника. Этот указатель позволяет обозреть всю информацию о каждом данном слове, сопоставить различные толкования и оценить, в какой мере они увязаны друг с другом.

xxx

Моделирование владения смыслом предполагает, как указывалось выше, наличие словаря и правил перевода. До сих пор мы рассматривали статьи сборника, относящиеся к первой части проблемы, - материалы к русско-семантическому словарю. Работы, посвященные правилам перевода, составляют меньшую по объему, но не менее интересную часть сборника. В отличие от "словарных" работ, работы "алгоритмические" не образуют системы. Каждая из них решает более или менее частный вопрос, связанный с составлением правил перевода. К тому же они относятся к разным "хронологическим эпохам"[10].

Наиболее общим является подход, развиваемый в работе А.К. Жолковского /2/. В статье описывается пробный алгоритм анализа предложения /перевода его с английского на семантический язык/. Для такого анализа предполагается имеющимся словарь, в котором каждой основе сопоставлена некоторая ситуационная форма - нульместная /для имен/, одноместная или двухместная. Роль предикатных символов играют элементарные смыслы, выделяемые в данной основе. Глубина формы может быть больше единицы, то есть место аргумента в форме может быть замещено другой формой. Например, семантическая запись слова convince /убеждать/ имеет вид: make (x, think (y, z)). Фактически у автора соответствующая словарная статья записана по другому:

P	1	2		
make		P	1	2
		think		

Однако, очевидна эквивалентность этого табличного способа записи формы скобочному. Хотя семантическая запись значений слов, представленная в этой статье, гораздо грубее семантической записи значений в рассмотренных выше словарных работах, она имеет то неоспоримое преимущество, что является последовательно формальной. Это и понятно, так как в противном случае используемый автором миниатюрный словарь не мог бы обслуживать алгоритма.

Семантическому анализу подлежит предложение с уже найденным деревом синтаксических зависимостей между словами. Анализ проводится в три этапа. Первые два этапа - построение семантической записи крупной единицы из семантических записей мелких посредством замещения переменных /пустых клеток таблицы/ одной формы другими формами. На первом этапе /морфологический анализ/ семантическая запись словоформы комбинируется указанным способом из семантической записи основы и семантической записи, которая приписана аффиксу. При этом используется информация, содержащаяся в словаре. На втором этапе семантические записи словоформ комбинируются в семантические записи словосочетаний, а последние - в семантическую запись предложения. Здесь новым источником информации служит дерево синтаксических зависимостей слов предложения. Третий этап - приведение полученной формы к каноническому /кратчайшему/ виду с использованием так называемых смысловых равенств. Простейший пример смыслового равенства - равносильность двойного отрицания утверждению.

Достоинства и недостатки описанного в статье алгоритма связаны с тем, что он носит пробный характер. С одной стороны, он рассчитан на перевод всего нескольких фраз и не может рассматриваться как рабочий алгоритм машинного перевода. С другой стороны, все этапы перевода от английского предложения до его канонической семантической записи предстают перед читателем в ясном и отчетливом виде.

Если работа А.К. Жолковского посвящена организации алгоритма автоматического перевода в целом, то в статье Ю.К. Щеглова рассматривается один частный вопрос - вопрос о логическом акценте или подчеркивании, связанный с начальными этапами анализа и конечными этапами синтеза. Важное понятие подчеркивания /логического акцента/, введенное в обращении в ранних семантических работах ЛМП, разъясняется в предисловии А.К. Жолковского I следующим образом: подчеркивание играет роль оператора, указывающего, в каком направлении следует развивать осмысление определенной ситуации. Изучение подчеркивания позволяет подойти с новой, более общей точки зрения к некоторым семантическим явлениям, особенно синонимии. Помимо описания синонимии отдельных слов /собственность - имущество, владелец-обладатель/ становится возможным описание синонимии целых высказываний /X - собственность У = У - владелец X, с точностью до подчеркивания/. Предполагается, что изучение подчеркивания даст ключ к пониманию способности говорящего употреблять при построении высказывания с заданным значением практически любые слова, относящиеся к данной типовой ситуации.

Предложения я написал письмо и письмо написал я, описывающие одну и ту же внеязыковую ситуацию, различны по смыслу. В первом предложении сообщением является тот факт, что мной было написано письмо, а не что-либо другое /подчеркнуто письмо/, а во втором предложении - тот факт, что автором письма был я, а не кто-либо другой /подчеркнуто я/. В первом случае грамматическое подлежащее совпадает с логическим субъектом /ЛС/, а грамматическое сказуемое - с логическим предикатом /ЛП/. Во втором случае грамматические подлежащие и сказуемые остаются прежними, но ЛС и ЛП изменяются /подчеркивается другая часть сообщения/. Поскольку при этом смысл сообщения изменяется, хороший алгоритм анализа должен сопоставить этим предложениям две различных семантических записи, а хороший алгоритм синтеза должен по таким записям построить предложения, в которых будет подчеркнута нужная часть сообщения. Поскольку ЛП выражается, в основном порядком слов /ср. Это я, Он на Поезде приехал/ и /в устной речи/ интонацией, вторая задача сводится к построению алгоритма, расставляющего слова в предложении и приписывающего им некоторую интонацию.

Для решения этой задачи вводятся следующие определения. 1/ ЛП предложения является та его часть, которая при трансформации в вопрос заменяется вопросительным словом /ср. Это я. Кто это? - ЛП - я/. В вопросительном предложении может быть свой ЛП, выделяемый теми же средствами /в данном случае - вопросом к вопросу/. 2/ Собственно логическим субъектом /$ЛС_с$/ называется та часть предложения, которая при трансформации этого предложения в вопросительное становится ЛП последнего. В русском языке $ЛС_с$ помещается во фразе до ЛП. 3/ Логическим субъектом называется любая часть предложения, не являющаяся ни ЛП, ни $ЛС_с$.

Алгоритм расстановки слов в предложении и приписывания интонации словам или группам слов работает после того, как целиком синтезированы синтаксическое дерево данного предложения и все его словоформы. Предполагается, что в дереве тем или иным способом указано, какую часть сообщения желательно выде-

лить в качестве ЛП. Сначала алгоритм расставляет группы слов, выражающие ЛП и $ЛС_с$ /или ЛС/. То, что желательно выделить в качестве ЛП, ставится в конец предложения; то, что желательно выделить в качестве $ЛС_с$ или ЛС, ставится в начало предложения. Оставшиеся нерасставленными группы расставляются по следующему общему правилу: группа первого обстоятельства - группа подлежащего - группа второго обстоятельства - группа сказуемого, внутри которой на первом месте стоит первое дополнение, затем глагол, затем второе и третье дополнения. К этим трем правилам добавляются, для случая устной речи, интонационные правила и связанные с ними перестановки.

Правила являются пробными и имеют чисто иллюстративный смысл. Однако, несмотря на то, что они далеко не охватывают всех возможных случаев, они обеспечивают синтез значительного числа предложений с разными ЛП и $ЛС_с$ /разными подчеркиваниями/ на основе одного и того же дерева. Так, члены предложения Комитет по премиям может ввести вас в свой состав могут быть расставлены двадцатью различными способами в зависимости от того, какая часть выделяется в нем как ЛП и какая - как $ЛС_с$.

Нам кажется, что приводимые в статье трансформационные правила идентификации ЛП и $ЛС_с$, содержательно весьма интересные, не могут быть формализованы, и поэтому для работы алгоритмов автоматического анализа информация об ЛП и $ЛС_с$ должна добываться другим путем. Алгоритмически значительную долю информации об ЛП и $ЛС_с$ можно извлечь из порядка слов переводимого предложения, расстановки знаков препинания, выбора предикатного слова /ср. выше пример с конверсными предикатами/. Однако, этих вопросов автор практически не рассматривает, и поэтому значение его статьи сводится, в основном, к постановке проблемы и указанию некоторых путей ее решения.

Особое место в сборнике занимает чрезвычайно интересная статья Ю.С. Мартемьянова /8/. Если у А.К. Жолковского и Ю.К. Щеглова контекстом, в пределах которого осуществляется переработка информации, является одно изолированное предложение, то для Ю.С. Мартемьянова контекстом оказывается весь текст целиком. В связи с этим возникает и более глубокий взгляд на понимание текста. Автор вводит представление о двух ступенях понимания. На первой ступени читатель осведомляется о внеязыковой ситуации /объектах и связывающих их отношениях/, являющейся непосредственным значением данного отрезка текста. На второй ступени читатель понимает место данной ситуации в более широком порядке явлений действительности, что дает ему возможность ориентироваться в этой ситуации /в частности, связывать ее с предшествующими, сопутствующими и последующими ситуациями/. Ставится задача смоделировать понимание читателем текста во втором смысле. Ниже мы попытаемся изложить предлагаемое Ю.С. Мартемьяновым решение этой задачи в том виде, как нам удалось ее понять.

Ход событий, описываемых текстом, представляется как последовательность состояний некоторой системы, элементы которой могут менять во времени свои свойства и пространственное расположение. Развитие действия определяется столкновением одушевленных элементов системы /героев/ друг с другом и с ее неодушевленными элементами и проявляется в поведении героев. Текущее поведение героя /или поступок/ на данном шаге развития действия является его ответом на ситуацию, представляющую собой предшествующее состояние системы. Этот ответ зависит от того, что герой знает об этой ситуации и каков его характер, и направлен на восстановление равновесного состояния его "души". Знания героя о ситуации складываются из образа последней, воспринимаемого им непо-

средственно, и его знаний о законах действительности. Понять текст в указанном выше втором смысле значит приписать "душе" героя такие знания о мире, такие способности восприятия наличной ситуации и такие правила реагирования на воспринимаемую ситуацию, чтобы, подав на вход этих правил предшествующую ситуацию, получить на выходе его поступок, описываемый текущим предложением текста.

Чтобы пояснить этот подход, рассмотрим в качестве примера следующий текст: "Обхаянному на лекции брюнету он пожертвовал даже ферзя. Брюнет пришел в ужас, но только страшным усилием воли заставил себя продолжать игру". В ситуации, описываемой первым из этих двух предложений, васюковский любитель шахмат непосредственно воспринимает только два момента: 1/ его противник - гроссмейстер и 2/ он пожертвовал ферзя. Он знает, что гроссмейстер должен играть хорошо, и что хороший игрок жертвует ферзя только будучи абсолютно уверенным в победе. Он знает, наконец, что в безнадежном положении следует сдаваться. Однако, свойства его характера таковы, что заставляют его продолжать игру. Так объясняется его поступок, описываемый в качестве текущего поведения во втором из наших предложений.

Формально "душа" героя описывается следующими документами: 1/ регистр образов, 2/ регистр состояний, 3/ таблица "понимания", 4/ алгоритм "понимания", 5/ алгоритм "реагирования", 6/ управляющий алгоритм. На каждом шаге развития в регистре состояний записано одно из 6 возможных состояний "души" /одно - равновесное, остальные - неравновесные/. Управляющий алгоритм опрашивает регистр состояний. Если состояние равновесное, герою предписывается продолжение поведения, обусловленного какими-то ранними ситуациями, для чего включается алгоритм "реагирования", вырабатывающий соответствующий поступок. Если состояние неравновесное, то алгоритму "реагирования" предлагается искать путь к восстановлению равновесия. В обоих случаях алгоритм "реагирования" выбирает поступки, справляясь с содержимым таблицы "понимания" и регистра образов. В таблице "понимания" отражены а/ знания героя в виде законов типа "Если А видит В, то А стремится понять В", б/ его характер в виде указания на его отношение к данному закону - положительное или отрицательное и в/ его способности в виде указания на возможность или необходимость приложения собственного усилия к выполнению этого закона. В регистре образов имеются места для образа восприятия, где всегда хранится описание наличной ситуации, и место для образа представления. Алгоритм "понимания", считывая образ восприятия, вырабатывает образ представления в соответствии с данными таблицы "понимания" и заполняет соответствующее место в регистре образов. Содержимым последнего, в свою очередь, определяется то значение, которое будет записано в следующий момент в регистр состояний.

Описание наличной ситуации, хранящееся в регистре образов, представляет собой, по замыслу автора, выражение на некоем символическом языке, усвоить который нам не удалось. Помимо этого языка Ю.С. Мартемьянов предлагает другой, табличный способ формального описания ситуации, изложение которого дает ему повод высказать несколько глубоких замечаний о приемах выражения содержания, используемых естественными языками. В частности, полное формальное задание ситуации /понимаемой как мгновенное состояние некоторой системы объектов/ исчерпывается перечнем элементов системы с указанием для каждого элемента, а также для каждой пары, тройки и т.д. элементов имеющихся у них в данный момент свойств. Естественно представлять ситуацию в виде совокупности пар вида А - В, где А - элемент / n-ка элементов/, а В - приписываемое ему

/им/ свойство. При этом одна ситуация от другой отличается лишь свойствами тех или других элементов. Но поскольку от момента к моменту изменяются свойства не всех элементов, то часть пар А - В в представлении текущей ситуации оказывается избыточной. Поэтому естественно, далее, сохранять в представлении ситуации только те пары, которые изменились против предыдущей ситуации. Именно так и устроено предложение естественного языка, в чем осуществляется присущее последнему стремление к краткости.

Богатство и смелость содержащихся в работе идей вызывает восхищение, но мы не ручаемся за верность их передачи и не чувствуем при этом никаких угрызений, поскольку статья написана чрезвычайно небрежно и исключает возможность полагаться на прямой смысл слов.

Кроме этих оригинальных статей в сборнике помещены два обзора: А.К. Жолковского о работах Э. Сэпира по структурной семантике и В.А. Матвеенко о семантических исследованиях группы С. Чекатто. Обзор А.К. Жолковского играет в сборнике особую роль. В нем впервые обращено внимание на те идеи и принципы Э. Сэпира, которые можно рассматривать как непосредственную основу современных семантических исследований.

Для этих исследований сохраняют значение следующие принципы и результаты Э. Сэпира. 1/ Выделение ряда элементарных и общих семантических признаков, свободных от идиоматичности и в силу этого годных для описания слов самых различных языков, и изображение значения слова в виде "пересказа" на языке этих признаков. В специальной работе о значениях слов, участвующих в выражении "отношения конечной точки", для таких пересказов используется семантический язык, содержащий, в частности, следующие признаки: а/ предикативность - непредикативность /слово кончается в предицирует отношение конечной точки, а слово к выражает его в качестве определения/; б/ конечная точка и точка отсчета; в/ движение /самостоятельное перемещение/ и двигание /перемещение объекта/ и т.п. 2/ Разграничение эксплицитных и имплицитных значений и учет последних в семантических определениях. Так, слову много приписывается значение сравнительной степени /"больше нормы"/, хотя внешне это значение никак не выражено. Данный принцип имеет и более общий смысл: поскольку семантические сходства и различия устанавливаются сразу для целой группы слов, то, хотя разные слова держат в фокусе разные части единой ситуации, каждое из них имплицитно обозначает и те элементы ситуации, выражению которых оно специально не посвящено. 3/ Установление семантических связей не между отдельными словами, а между целыми высказываниями. 4/ Открытие того факта, что в значениях слов, обозначающих градуирование /а может быть, и других слов/. Объективный признак степени /ср. больше, меньше/ переплетается с субъективным значением оценки, зависящим от желательности или нежелательности увеличения или уменьшения предмета мысли. Та или иная комбинация объективных и субъективных компонентов значения жестко определяет выбор слов, которые могут быть использованы для ее выражения. Так, если человеку осталось ждать желательного события три с половиной дня, и он считает этот срок небольшим, то, выражаясь при помощи числительного три и градуирующего слова больше, он не может сказать Мне осталось ждать больше трех дней; в этом случае ему следует сказать Мне осталось ждать чуть-чуть больше трех дней.

ххх

Ниже мы постараемся выделить и оценить основные достоинства и недостатки рассматриваемого сборника в целом и сделать выводы о его месте в разработке современных вопросов лингвистической семантики.

Разрабатываемая лабораторией семантическая модель должна состоять, как было сказано выше, из двух частей: естественно-семантических и семантико-естественных словарей и алгоритмов перевода. В традиционной семантике нет ничего, что соответствовало бы алгоритмам перевода, так как в число ее задач никогда не включалась задача смоделировать языковое поведение человека, владеющего значением слов. Поэтому мы можем сравнивать с обычными семантическими /лексикографическими/ работами и представлениями лишь первую часть модели, а именно - материалы к русско-семантическому словарю.

Представим себе обычный толковый или двуязычный словарь. Каждая статья в таком словаре имеет две части: определяемое слово и его толкование с помощью некоторых определяющих слов. И выбор множества определяемых слов и выбор множества определяющих слов, участвующих в толкованиях, выдвигают несколько очень важных вопросов. Мы обсудим сначала вопросы, связанные с выбором множества определяемых слов.

С точки зрения авторов сборника, далеко не все имеющиеся в языке /определяемые/ слова интересны для лингвиста; строгое определение таких, например, слов, как иволга, фламинго, альбатрос возможно на языке зоологии, но не на языке лингвистики. С этой точки зрения каждый толковый словарь является, по существу, плохим энциклопедическим словарем ровно в той мере, в какой он содержит толкования слов, для определения которых лингвист не располагает необходимыми понятиями.

Русско-семантический словарь строится иным образом - с учетом различия между словами типа иволга, фламинго, альбатрос и словами типа воля, законченность, желательность, причина, результат и т.п. Слова первого типа, обозначающие предметы, в русско-семантическом словаре описываются достаточно поверхностно. В крайнем случае их определением является порядковый номер в словаре. Семантический код слов второго типа, обозначающих отношения, гораздо более богат. Именно эти слова представляют наибольший интерес для лингвиста. по предположению авторов, это - те слова, значения которых, хотя бы в некоторых языках, являются грамматическими, то есть выражаются в обязательном порядке и поэтому входят в правила кодирования сообщений на данном языке. Как показывает рассмотренный выше материал, способность именно таких слов превращаться друг в друга лежит в основе семантических преобразований. Интересно, что слова, обозначающие предметы, при всех этих преобразованиях сохраняются.

Вход словарной статьи в русско-семантическом словаре отличается от входа словарной статьи в толковом или двуязычном словаре еще в некоторых отношениях. Прежде всего, как было сказано выше, входом словарной статьи в русско-семантическом словаре является не слово, а ситуационная форма. Именно этот принцип позволяет по-новому поставить вопрос о синонимии: синонимичными могут быть и такие ситуационные формы, которые не содержат синонимов в обычном смысле слова. Благодаря этому обнаруживаются такие семантические связи между словами /например, конверсия/, которые до сих пор либо вовсе не замечались, либо не были изучены систематически.

Выбрать множество определяющих слов значит выбрать метаязык описания. Проблема метаязыка описания никогда не ставилась в отчетливой форме в традиционной лексикографии. По существу, лексикограф подходил и к определяемым, и к определяющим словам как словам одного и того же языка. В результате появлялись неизбежные в словарях старого типа тавтологические определения, например, определение помощи через поддержку и поддержки через помощь или

определение прилагательного значительный как большой по размерам, а прилагательного большой как значительный по размерам /С.И. Ожегов/.

В русско-семантическом словаре, разрабатываемом Лабораторией МП, такой метаязык есть. Это язык элементарных смыслов. В рассматриваемых материалах любое определение либо может быть развернуто в запись на языке элементарных смыслов, либо уже записано в такой форме. Яркой иллюстрацией преимуществ, которые дает исследователю четкая постановка вопроса о семантическом метаязыке, является элементарное метаязыковое понятие "нормы", последовательно проводимое во всем сборнике с прекрасными результатами. Как это ни курьезно, понятие нормы, при всей его естественности, совершенно не используется в обычных толковых словарях и, таким образом, должно считаться лексикографическим открытием ЛМП. Благодаря ему авторы сборника успешно справляются с определениями весьма многочисленных в языке слов типа большой, значительный, огромный, крайний, чрезвычайный и других подобных, которые для авторов обычных словарей представляют непреодолимые трудности /см. выше примеры тавтологических определений слов большой и значительный/.

Отметим, наконец, что русско-семантический словарь обладает интересными свойствами словарей двух типов: толковых и идеологических. В толковых словарях слова семантически не упорядочены, но каждое слово имеет определение. В идеологических словарях слова не имеют определений, но семантически упорядочены. В русско-семантическом словаре налицо и семантическая упорядоченность на входе, и толкование на выходе [11].

Семантические работы ЛМП имеют две "ипостаси": во-первых, это - материалы к русско-семантическому словарю и алгоритмы перевода, и в качестве таковых они могут быть сопоставлены с аналогичными практическими работами традиционного характера, например, словарями. Это мы и попытались сделать выше. Во-вторых, они содержат изложение определенной теоретической концепции и поэтому могут быть сопоставлены с обычными теоретическими исследованиями в области лингвистической семантики. Классическая лингвистическая семантика, как она изложена, например, С.У. Ульманом [12], представляет собой систематику семантических явлений, имеющих место в пределах слова или класса слов. К ним относятся многозначность и омонимия, синонимия и антонимия, значение и употребление, типы лексических значений /связанные и свободные, обусловленные и необусловленные/ и т.п. Поэтому в рамках классической семантики не мог быть и не был поставлен гораздо более глубокий и серьезный вопрос о тех семантических механизмах языка, которые делают возможным осмысленное речевое поведение человека. Как мы видели, для того, чтобы научно поставить этот вопрос, оказалось необходимым перейти от описания значений отдельных слов к описанию значений семантически более законченных единиц - ситуационных форм. Этот шаг и был сделан в работах ЛМП.

Семантические исследования ЛМП выгодно отличаются не только от традиционных, но и от многих современных работ в этой области. Оставляя в стороне психолингвистические исследования Ч. Осгуда и его группы [13], которые не имеют прямого отношения к лингвистической семантике, как она понимается в рассматриваемом сборнике, и мало понятные, хотя, может быть, и не лишенные смысла семантические исследования С. Чеккато и его коллег, скажем несколько слов о работах Кэмбриджского лингвистического кружка, наиболее близким по духу к семантическим исследованиям ЛМП.

Хотя Кэмбриджский лингвистический кружок разработал формальный аппарат теории несравненно более тщательно и подробно, чем ЛМП, кэмбриджцы далеко не

достигают той глубины семантического анализа, которая отличает лучшие работы ЛМП. Семантический словарь используемый Кэмбриджским лингвистическим кружком, содержит 100 слов /минимальных элементов системы/; в число этих элементов входят, с одной стороны, некоторые весьма неэлементарные слова /считать, мечтать, догадываться, покупать, продавать и т.п./, а с другой стороны - ряд совершенно конкретных слов /животное, растение, мягкий, мокрый и др./. Семантический код любого семантически разложимого слова составляется непосредственно из этих элементов, в то время как определение предикатного символа в словаре ЛМП строится ступенчато; как мы помним, в толковании слова в словаре ЛМП помимо весьма простых и общих элементарных смыслов, могут участвовать и практически всегда участвуют промежуточные и любые уже определенные понятия. Такой порядок весьма близко соответствует процессу образования абстракций при обучении человека значениям слов, когда он переходит от менее абстрактных к более абстрактным словам.

Здесь еще раз проявляется одна из основных особенностей сборника - пристальное внимание к наиболее существенным моментам языкового процесса. Указанная особенность, а также глубина и смелость с которой авторы подходят к постановке и решению многих труднейших задач лингвистической семантики, делают появление сборника, несмотря на его очевидные многочисленные недостатки, выдающимся научным событием.

[1] См. А.К. Жолковский, Н.Н. Леонтьева, Ю.С. Мартемьянов. О принципиальном использовании смысла при машинном переводе. Сб. "Машинный перевод. Труды Института ТМ и ВТ АН СССР", вып. 2, 1961; Н.Н. Леонтьева.Модель синтеза русской фразы на основе семантической записи. "Доклады на конференции по обработке информации, машинному переводу и автоматическому чтению текстов", М., 1961; бюл. "Машинный перевод и прикладная лингвистика", вып. 8, М., 1964. В данной рецензии мы рассматриваем, в основном, материалы последнего сборника, в котором идеи группы ЛМП нашли наиболее полное выражение.

[2] Пример из статьи "О принципиальном использовании смысла при машинном переводе", см. предыдущую сноску.

[3] Сб. "Машинный перевод и прикладная лингвистика", М. 1964. В дальнейшем для ссылок на статьи этого сборника мы будем использовать цифры, заключенные в скобки; они соответствуют номерам статей сборника.

[4] В рассматриваемой книге всюду, где идет речь о формах, употребляется термин "предикат". Это некорректно, поскольку предикат - не форма, а присоединяемая к форме функция. Кроме того, поскольку здесь имеются в виду не высказывательные, а ситуационные формы, в указанном словоупотреблении содержится еще одна неточность. Ее, правда, можно было бы избежать, если обобщить понятие предиката и на функцию, присоединяемую к ситуационной форме.

[5] Эти сведения образуют в совокупности то, что авторы называют "наивной физикой" соответствующего отрезка действительности.

[6] Термин "конверсия" заимствован из общей алгебры. Отношение R' называется обратным /конверсным/ к отношению R, если из того, что элементы $\langle a, b \rangle$

находятся в отношении R, следует, что элементы $\langle b,a\rangle$ находятся в отношении R', и наоборот. Примерами конверсных слов, /рассматриваемых нами как имена отношений/ являются слова типа строить - строиться, раньше - позже, отец - сын, сообщать - узнавать - сообщаться /ср. он сообщает мне новости - я узнаю новости от него - новости сообщаются мне им /. В последнем примере глаголы обозначают трехместные отношения, между тем, как данное выше определение имеет силу лишь для двухместных отношений. Один из авторов данной статьи /К.И. Бабицкий/ дал следующее обобщение понятия конверсии для n-местных отношений: отношение Q конверсно отношению R, если существует взаимно-однозначное отображение γ множества Q на множество R, представляющее собой определенную подстановку.

[7] Здесь и ниже выражения "активная /пассивная/ сила" употребляются как удобный синоним для выражения "сила в активном /пассивном/ состоянии".

[8] В общем случае это различие не совпадает с различием отчуждаемых и неотчуждаемых ценностей. Пример: право наследования - идеальная но отчуждаемая ценность.

[9] Роль стандартных перифраз, разумеется, выходит за рамки задачи описания только этого конкретного предлога.

[10] В частности, статья А.К. Жолковского была написана в 1961 году.

[11] Сказанное справедливо лишь при условии, что алфавитный указатель, придаваемый обычно идеологическому словарю, не рассматривается как его существенная составная часть.

[12] S. Ullmann. The Principles of Semantics. Glasgow, 1957.

[13] Ch. Osgood, L. Suci, H. Tannenbaum. The Measurement of Meaning Urbana,1957.

ОПЫТ АВТОМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ НА ЯЗЫКЕ ЭСПЕРАНТО

И. Ботош

Статья посвящается 51 Всемирному конгрессу эсперанто в Будапеште /1966 г./

В Вычислительном центре при Венгерской Академии наук вырабатывается автоматический анализ текстов на русском языке. Употребляемый при этом метод выделения категорий и составления правил для синтаксического анализа 1 применяется в данной статье к языку эсперанто. /В основе метода лежит алгоритм В. Домелки, публикуемый в этом же томе Computational Linguistics. Метод применен Д. Варгой к анализу русского языка - при более сложных условиях естественных языков. Описание анализа русского языка с помощью указанного алгоритма публикуется в ближайшем будущем./ Кроме изменений, объясняемых спецификой искусственного международного языка эсперанто, делается здесь попытка разбить правила в блоки, применяемые при синтаксическом анализе один за другим, в целях уменьшения количества "тупиков" и вынужденных повторений известных процессов анализа.

Для показа системы возьмем десять предложений из статьи Шандора Сатмари "Kion scii pri Hungario por kongresanoj?" (Hungara Vivo, N° 1. 1966) и представим те части словаря /списка морфем/, списка категорий /символов/ и блоков правил, которые используются для анализа данных десяти предложений.

Вот предложения /с дословным по возможности переводом на русский язык/

1. Kion scii pri Hungario por kongresanoj? 'Что знать о Венгрии участникам Конгресса?'

2. La 51-a Universala Kongreso okazos Budapeŝte. 51-ый Всемирный конгресс состоится в Будапеште.'

3. Je ĉi tiu okazo ni intencas doni kelkajn indikojn, kiujn niaj gastoj devas scii pri la lando. 'По этому случаю мы намерены дать несколько данных, которые наши гости должны знать о стране.'

4. Hungario situas apud la riveroj Danubo kaj Tibisko, sur la limo de Mez- kaj Orienteuropo. 'Венгрия расположена у рек Дуная и Тисы, на границе Средней и Восточной Европы.'

5. La landon limas Austrio, Ceĥoslovakio, Sovetio, Rumanio kaj Jugoslavio. 'Со страной граничат Австрия, Чехословакия, Советский Союз, Румыния и Югославия.'

6. La popoldenseco estas 109/km^2. 'Плотность населения - 109/км^2.'

7. La klimato estas kontinentala. 'Климат континентальный.'

8. La meza temperaturo estas 10.9 C°. 'Средняя температура - 10,9°С.

9. Plej granda parto de la lando (67%) estas fekunda ebenaĵo, iama marfundo. 'Наибольшая часть страны /67%/ - плодородная низменность, бывшее дно моря.'

10. Ce la norda limo de la Granda Ebenaĵo situas la ĉefurbo Budapest, "la reĝino de la Danubo". 'У северной границы Большой Низменности расположена столица Будапешт, "королева Дуная".'

Список морфем

Порядок знаков /алфавит/: Ø /пробел/ a b c ĉ d e f g ĝ h ĥ i j ĵ k l m n o p r s ŝ t u u v z . ? , () "

Морфема	№ таблицы окончаний	Символ
apudø		F01
austr	4	###
budapestø		R05
budapeŝt	2	##0
ĉeø		F04
ĉef	2	##0
ĉeĥoslovak	4	###
ĉiø		B01
danub	2	##5
deø		F02
dens	2	##0
dev	2	##9
don	2	##3
eben	2	##0
est	2	##7
europ	2	##0
fekund	2	##0
fund	3	###
gast	2	##0
grand	2	##0
hungar	4	###
iam	6	###
indik	2	##2
intenc	2	##9
jeø		F00
jugoslav	4	###
kajø		J00
kelk	5	###
kio	1	P#5
kiu	1	P#4
klimat	2	##0
kongres	2	##0
kontinental	2	##0
laø		A40
land	2	##0
lim	7	###
mar	2	##0
mez	3	###
niø		P11
nia	1	A#1
nord	3	###
okaz	3	###
orient	7	###

Морфема	№ таблицы окончаний	Символ
part	2	##0
plejø		B03
popol	7	###
porø		F03
priø		F09
reĝ	2	##0
river	2	##0
ruman	4	####
sci	2	##2
situ	2	##0
sovet	4	###
surø		F01
temperatur	2	##0
tibisk	2	##5
tiu	1	P#6
universal	2	##0
urb	2	##0
.ø		Z00
?ø		Z02
,ø		Z10
(ø		Z13
)ø		Z14
"ø		Z32
/арабская цифра + черточка/	2	##0
/формула/		R01

Вторая графа "списка морфем" № таблицы указывает на номер нижеследующих таблиц окончаний и суффиксов.

Таблицы суффиксов и окончаний

№ таблицы	Морфема /или сочетание морфем/	Ссылка на другую таблицу	Символ
1	ø		#0#
	jø		#1#
	nø		#2#
	jnø		#3#
2	iø		V0#
	asø		V1#
	isø		V1#
	osø		V1#

№ таблицы	Морфема /или сочетание морфем/	Ссылка на другую таблицу	Символ
2	usø		V2ж
	uø		V3ж
	o	1	Nжж
	ano	1	Nжж
	ajo	1	Nжж
	ino	1	Nжж
	eco	1	Nжж
	a	1	Aжж
	e	1	Bжж
3	iø		V00
	asø		V10
	isø		V10
	osø		V10
	usø		V20
	uø		V30
	o	1	Nж0
	ano	1	Nж0
	ajo	1	Nж0
	eco	1	Nж0
	a	1	Aж0
	e	1	Bж4
	-ø		R00
4	o	1	Nж0
	ajo	1	Nж0
	eco	1	Nж0
	io	1	Nж1
	ino	1	Nж0
	a	1	Aж0
	e	1	Bж0
5	a	1	Aж0
	eø		B05
6	ø		B07
	a	1	Aж0
7	iø		V03
	asø		V13
	isø		V13
	osø		V13
	usø		V23
	uø		V33
	o	1	Nж0
	ano	1	Nж0
	ajo	1	Nж0

№ таблицы	Морфема /или сочетание морфем/	Ссылка на другую таблицу	Символ
7	eco	1	N*0
	a	1	A*0
	e	1	B*4
	-ø		R00

Специальное правило для сложных слов

Если устанавливается наличие в слове рядом нескольких корневых морфем /т.е. морфем, перечисленных в "списке морфем" и оканчивающихся не на ø/, то принимается во внимание информация /№ таблицы окончаний и символ/ только последней из этих морфем, и на место последнего /третьего/ знака символа ставится цифра 6.

Установление символа

В целях анализа слово заменяется символом, указывающим принадлежность к определенному структурному классу слов. Механизм установления символа следующий: слово сравнивается с морфемами, помещенными в "списке морфем". Последней буквой слова считается находящийся за ним пробел. Когда в "списке морфем" найдена самая длинная морфема, совпадающая с началом искомого слова, в "таблице суффиксов и окончаний", номер которой указан во второй графе, отыскивается продолжение слова. В случае неудачи продолжение слова отыскивается в "списке морфем" /сложные слова, см. выше/. В "таблицах суффиксов и окончаний" при элементах, оканчивающихся не на ø, находится "ссылка на другую таблицу", где и должен отыскиваться конец слова. Символ данного слова /данной словоформы/ устанавливается путем сочетания элементов /знаков/ символа, находящихся в последней графе как "списка морфем", так и "таблиц суффиксов и окончаний". /Найденный знак символа ставится при этом на место звездочки./ Если морфема в "списке" оканчивается на ø, номер таблицы при ней не указан, а символ целиком помещен в "списке". Если в "списке" графа символа пуста, он целиком зависит от суффиксов и окончаний.

Примеры установления символа:

1. Слово kion/kionø — По "списку":
   kio — |1|P*5|
   По таблице № 1:
   nø — | |*2*|
   Итак символ, заменяющий словоформу: P25
2. Слово Hungario. — По "списку":
   hungar — |4|***|
   По таблице № 4:
   io — |1| |
   По таблице № 1:
   ø — | |*0*|
   Результат: N01
3. Слово scii — По "списку":
   sci — |2|**2|
   По таблице № 2:
   iø — | |V0*|

Итак символ: V02

4. Слово Orienteuropo. По списку:

orient	7	жжж

В таблице № 7 продолжение слова не содержится.

/Есть: e 1 Bж4 , но в таблице № 1 нет дальнейшего продолжения./

Поэтому следует вернуться к "списку":

europ	2	жж0

По "специальному правилу для сложных слов":

orienteurop	2	жж6

По таблице № 2:

o	1	Nжж

По таблице № 1:

ø		жОж

В результате: N06

Значение символов

Каждый символ состоит из трех знаков, первый из которых является буквой латинского алфавита, остальные два - арабскими цифрами. Символ может быть исходным /соответствующим терминальному в генеративной грамматике/, неисходным, или смешанным /могущим обозначать как словоформу, так и сочетание слов/.

Значение первого знака символа

A - прилагательное, числительное, притяжательное местоимение, артикль , определительное придаточное предложение
B - наречие, частица
F - предлог, предложный оборот
J - союз
N - существительное /грамматически оформленное/
P - местоимение /не входят сюда местоименные прилагательные/
R - слово, грамматически неоформленное /формула, неоформленное с точки врения языка эсперанто географическое название, корень + черточка
S - предложение
V - глагол
Z - знак препинания

Значение второго знака символа

0 - после A, N, P - форма общего /именительного и совпадающего с ним предложного/ падежа единственного числа /нулевое окончание/
- после B - наречие, отвечающее на вопрос "куда" /нулевое окончание/, частица
- после F - предлог
- после J - сочинительный союз
- после S - предложение с заключительным знаком препинания
- после V - инфинитив /-i/
- после Z - заключительный знак препинания

1 - после A, N, P - форма общего падежа множественного числа или нескольких сочиненных имен
- после S - предложения с двумя главными членами
- после V - изъявительное наклонение

- после Z - запятая, скобки

2 - после A, N, P - форма винительного падежа единственного числа /-n/
- после B - наречие, отвечающее на вопрос "куда"
- после V - условное наклонение

3 - после A, N, P - форма винительного падежа множественного числа или нескольких сочиненных имен
- после F - предложный оборот
- после V - повелительное наклонение

4 - после A - артикль, числительное
- после V - сказуемое, могущее употребляться только при подлежащем единственного числа

5 - после A - определительное придаточное предложение

Значение третьего знака символа

0 - после A0, A1, A2, A3, B0, B2, N0, N1, N2, N3, V0, V1, V2, V3 - слово без специального управления
- после A4 - артикль
- после A5 - придаточное предложение, являющееся определением имени в форме единственного числа
- после F0 - предлог je
- после F3 - предложный оборот с предлогом je
- после J0 - союз kaj
- после P0, P2 - местоимение vi
- после R0 - корень + черточка
- после Z0 - точка
- после Z1 -запятая

1 - после A0, A1, A2, A3 - притяжательное местоимение
- после A4 - количественное числительное unu
- после A5 - придаточное предложение, являющееся определением имени в форме множественного числа
- после B0 - частица ĉi
- после F0 - предлог места с обоими /общим и винительным/ падежами
- после F3 - F01 + имя /именное сочетание/ в форма общего падежа
- после N0, N1, N2, N3 - название страны
- после P0, P2 - местоимение mi
- после P1, P3 - местоимение ni
- после R0 - формула

2 - после A4 - количественное числительное, за исключением unu
- после F0 - предлог de
- после F3 - предложный оборот с предлогом de
- после N0, N1, N2, N3, V0, V1, V2, V3 - слово, управляющее винительным падежом и предлогом pri
- после P0, P2 - местоимение ci
- после R0 - формула в скобках
- после Z0 - вопросительный знак

3 - после A0, A1, A2, A3 - прилагательное с частицей plej /превосходная степень/
- после B0 - частица plej
- после F0 - предлог por
- после F3 - оборот с предлогом por

- после N0, N1, N2, N3 - частица plej + прилагательное + существительное
- после P0, P2 - местоимения li, ŝi, ĝi
- после P1, P3 - местоимение ili
- после V0, V1, V2, V3 - глагол, управляющий винительным падежом
- после Z1 - левая скобка

4 - после B0, B2 - наречие, управляющее предлогом de
- после F0 - предлог места с общим падежом
- после F3 - F04 + имя /именное сочетание/
- после N0, N1, N2, N3 - определенное существительное /определенное именное сочетание/
- после P0, P1, P2, P3 - местоимение kiu
- после Z1 - правая скобка

5 - после B0 - наречие, управляющее предлогом da
- после N0, N1, N2, N3 - название города или реки, грамматически оформленное
- после P0, P2 - местоимение kio
- после R0 - географическое название, грамматически неоформленное
- после V0 - местоимение kion + инфинитив переходного глагола

6 - после A0, A1, A2, A3, B0, B1, N0, N1, N3, N3, V0, V1, V2, V3 - сложное слово
- после P0, P1, P2, P3 - местоимение tiu

7 - после B0 - наречие iam
- после A0, A1, A2, A3, N0, N1, N2, N3, V0, V1, V2, V3 - словоформа от корня est- /esti = быть/
- после P0, P2 - местоимение tio

8 - слово или ряд слов в кавычках

9 - после F0 - предлог pri
- после F3 - оборот с предлогом pri
- после A0, A1, A2, A3, B0, N0, N1, N2, N3, V0, V1, V2, V3 - слово, управляющее инфинитивом

Список символов

В нижеследующем списке символов, используемых при анализе перечисленных выше десяти предложений, в графе 1 называется символ, в графе 2 перечисляются словоформы, обозначаемые данным символом, в графе 3а указаны те ряды символов, которые заменяются при анализе названным в графе 1 символом, и, наконец, в графе 3б отмечен номер блока правил, где помещено правило, в силу которого указанное в графе 3а сочетание символов заменяется символом, указанным в графе 1.

Если графа 3а /следовательно, и 3б/ пуста, то указанный в графе 1 символ является исходным.

Если графа 2 пуста, то указанный в графе 1 символ является неисходным.

Если ни одна из граф не пуста, то символ является смешанным.

1	2	3а	3б
A00	/арабская цифра/ - a universala kontinentala meza granda fekunda iama norda		
A03		B03 A00	1
A11	niaj		
A30	kelkajn		
A40	la		
A51		P34 N10 V12	4
B00	budapeŝte		
B01	ĉi		
B03	plej		
F00	je		
F01	apud sur		
F02	de		
F03	por		
F04	ĉe		
F09	pri		
F30		F00 N04	3
F31		F01 N04 F01 N14 F31 Z10 F31	3 3 3
F32		F02 N06 F02 N04	3 3
F33		F03 N10	3
F34		F04 N04	3
F39		F09 N01 F09 N04	3 3
J00	kaj		
N00	kongreso okazo lando	A00 N00 A00 N06	2 2

1	2	3a	3b
N00	klimato temperaturo parto ebenaĵo limo reĝino		
N01	hungario austrio ĉeĥoslovakio sovetio rumanio jugoslavio		
N03		A03 N00	2
N04		A40 N00 P06 N00 A40 N06 P06 N06 A40 N05 P06 N05 N04 F32 N03 F32 N04 R05 N03 R05 N04 R02 N04 Z10 N08	2 2 2 2 2 2 3 3 3 3 4 4
N05	danubo tibisko		
N06	orienteuropo popoldenseco marfundo ĉefurbo	R00 J00 N06	2
N08		Z32 N04 Z32	3
N10	kongresanoj gastoj riveroj	A11 N10 N00 Z10 N00	2 3
N11		N01 J00 N01 N01 Z10 N11 N01 J00 N11 N01 Z10 N01	3 3 3 3
N14		A40 N10 N14 N15	2 3

1	2	3a	3б
N15		N05 J00 N05	3
N20	landon		
N24		A40 N20	2
N30		A30 N32	2
N32	indikojn		
N34		N30 Z10 A51	4
P06	tiu	B01 P06	1
P11	ni		
P25	kion		
P34	kiujn		
R00	mez-		
R01	/формула/		
R02		Z13 R01 Z14	2
R05	budapest		
S00		S10 Z00	5
		V05 Z02	5
S10		N04 V10	4
		P11 V10	4
		F30 S10	4
		N01 V10	4
		V10 N11	4
		N04 V40	4
		V10 N04	4
V00		V03 N34	4
V02	scii	V02 F39	3
V03	doni		
V05		P25 V02	4
		V05 F33	4
V10	okazos	V10 B00	3
	situas	V19 V00	4
		V10 F31	3
		N24 V13	4
		V17 N10	3
		F34 V10	3
V12		V19 V02	3
V13	limas		
V17	estas		

1	2	3а	3б
V19	intencas devas		
V40		V17 A00 V17 R01	3 3
Z00	.		
Z02	?		
Z10	,		
Z13	(		
Z14	)		
Z32	"		

Правила синтаксического анализа

В первой вертикальной графе нижеследующих таблиц названы символы, могущие быть составными частями тех соединений символов, которые осуществляются по данному блоку правил. В первой горизонтальной графе названы результаты применения правил, т.е. символы, заменяющие собой ряд, состоящий из тех символов, в горизонтальной графе которых стоит цифра 1. Например, первое правило первого блока следует понимать так: B03 + A00 заменяется символом A03. Если в одной и той же вертикальной графе находится несколько цифр "1" одна под другой, это значит, что любой из отмеченных символов может играть роль данного элемента данного соединения символов. Так, например, по первому правилу блока № 2 вторым элементом соединения символов может быть как N00, так и N05 или N06.

Вертикальные черты обозначают границы отдельных правил. Они одновременно указывают, из скольких /двух или трех/ элементов состоит соединение символов по данному правилу.

Блок правил № 1

	A03	P06
A00	1	
B01		1
B03	1	
P06		1

Блок правил № 2

	N00	N03	N04	N06	N10	N14	N24	N30	R02
A00	1								
A03		1							
A11					1				
A30								1	
A40			1			1	1		
J00				1					
N00	1	1	1						
N05	1		1						
N06	1	1	1	1					
N10					1	1			
N20							1		
N32								1	
P06			1						
R00				1					
R01									1
Z13									1 1
Z14									1

Блок правил № 3

	F30	F31	F31	F32	F33	F34	F39	N04	N08	N10	N11	N14	N15	V02	V10	V10	V10	V40
A00																		1
B00														1	1	1		
F00	1																	
F01		1																
F02				1														
F03					1													
F04						1												
F09							1											
F31			1	1										1	1	1	1	
F32								1				1						
F34															1	1	1	
F39														1				
J00			1							1	1		1					
N00	1	1		1	1	1	1	1		1	1				1			
N01		1		1	1		1	1			1							1
N03	1	1		1	1	1	1	1									1	
N04	1	1		1	1	1	1	1									1	
N05		1		1	1	1	1	1					1					1
N06	1	1		1	1	1	1	1		1	1						1	
N10	1	1		1	1	1	1			1	1	1					1	
N11		1		1	1		1				1	1						
N14	1	1		1	1	1	1					1					1	
N15		1		1	1	1	1					1	1					
R01				1	1		1	1										1
R05		1		1	1	1	1	1										1
V02														1				
V10															1	1		
V17																	1	1
Z10			1							1	1		1					
Z32									1	1								

Блок правил № 4

	A51	N04	N04	N34	S10	S10	S10	S10	V00	V05	V05	V10	V10	V12
A51				1										
F30							1							
F31							1							
F33							1			1				
F34							1							
N00	1				1	1	1							
N01	1				1	1	1							
N03	1				1	1	1							
N04	1	1	1		1	1	1							
N05	1				1	1	1							
N06	1				1	1	1							
N08	1		1		1	1	1							
N10	1				1	1								
N11	1				1	1								
N14	1				1	1								
N20									1				1	
N24									1				1	
N30				1					1				1	
N32				1					1				1	
N34				1					1				1	
P11	1				1	1								
P25										1				
P34	1													
R02		1												
S10								1						
V00												1		
V02											1			1
V03									1					
V05											1			
V10					1	1								
V12	1													
V13	1													1
V19												1		1
V40							1							
Z10			1	1										

Блок правил № 5

	S00	S00
S10	1	
V05		1
Z00	1	
Z02		1

Ход анализа

1. Предложение заменяется рядом символов /см. выше "Установление символа"/. Предложение № 1, например, заменяется следующим рядом символов:

P25 V02 F09 N01 F03 N10 Z02

Соответствия: P25 - kion 'что' /винительный падеж/
V02 - scii 'знать'
F09 - pri 'о'
N01 - Hungario 'Венгрия'
F03 - por 'для'
N10 - kongresanoj 'участники конгресса'
N02 - ?

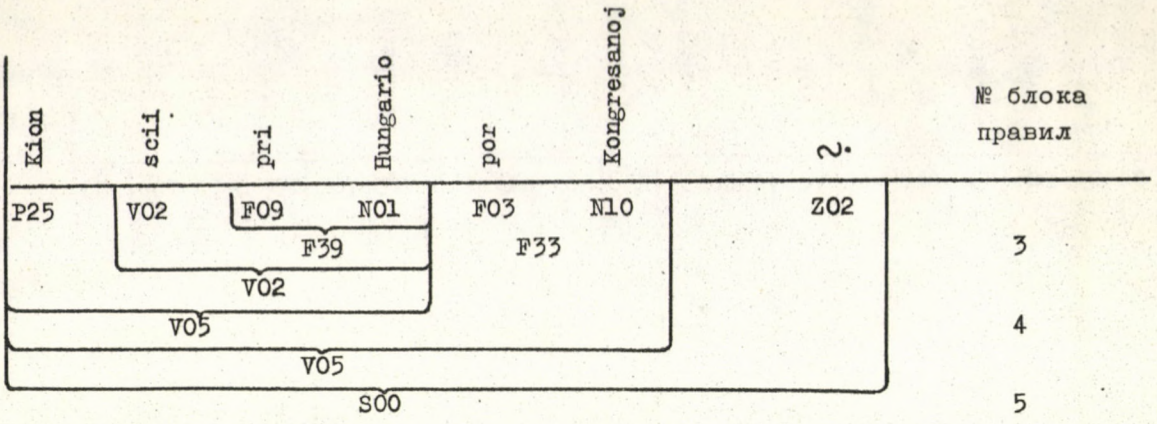
2. Анализ предложений на эсперанто происходит справа налево. Сперва правый крайний и стоящий перед ним символы /в нашем примере Z02 и N10/ сопоставляются с символами в первой вертикальной графе таблицы "Блока правил № 1". Поскольку совпадение в данном случае не обнаруживается, отыскивается там же символ второй - N10 справа /и стоящий перед ним F03 /, потом третий- F03 /и N01/, и т.д., до левых крайних символов - V02 и P25. Потом следует блок № 2. В блоке правил № 2 встречается, правда, символ N10, но не в сочетании налево с F03 а, по одному правилу, с A11, по другому - с A40. Первое соединение символов может осуществляться лишь по блоку № 3, где в вертикальной графе с головкой "F33" фигурируют: направо несколько символов, среди них и N10, а налево F03. В результате получается:

P25 V02 F09 N01 $\underbrace{\text{F03 N10}}_{\text{F33}}$ Z02

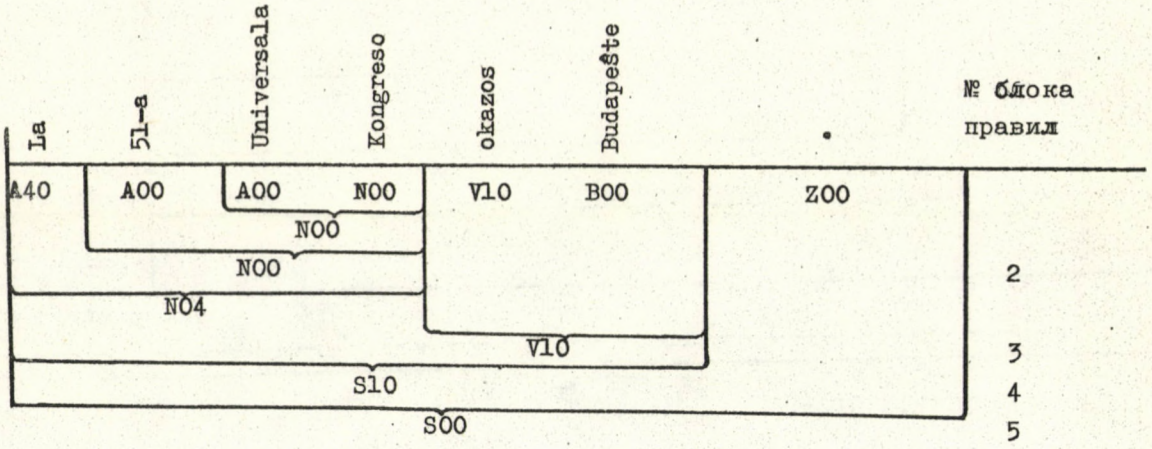
3. После каждой замены процесс повторяется /по тому же блоку правил/ снова с правого края ряда символов. /Новый ряд символов: P25 V02 F09 N01 F33 Z02/. Все еще по блоку правил № 3 устанавливается соединение символов F09 + N01, которое должно быть заменено символом F39 /Результат: P25 V02 F39 F33 Z02/. Потом соединяются V02 + F39 в символ V02 /P25 V02 F33 Z02/. Применяя правила блока № 4, получеатся ряд V05 F33 Z02, потом ряд V05 Z02, который, наконец, по одному из правил блока № 5, заменяется символом S00.

Показываем анализ наших десяти предложений:

Предложение № 1:

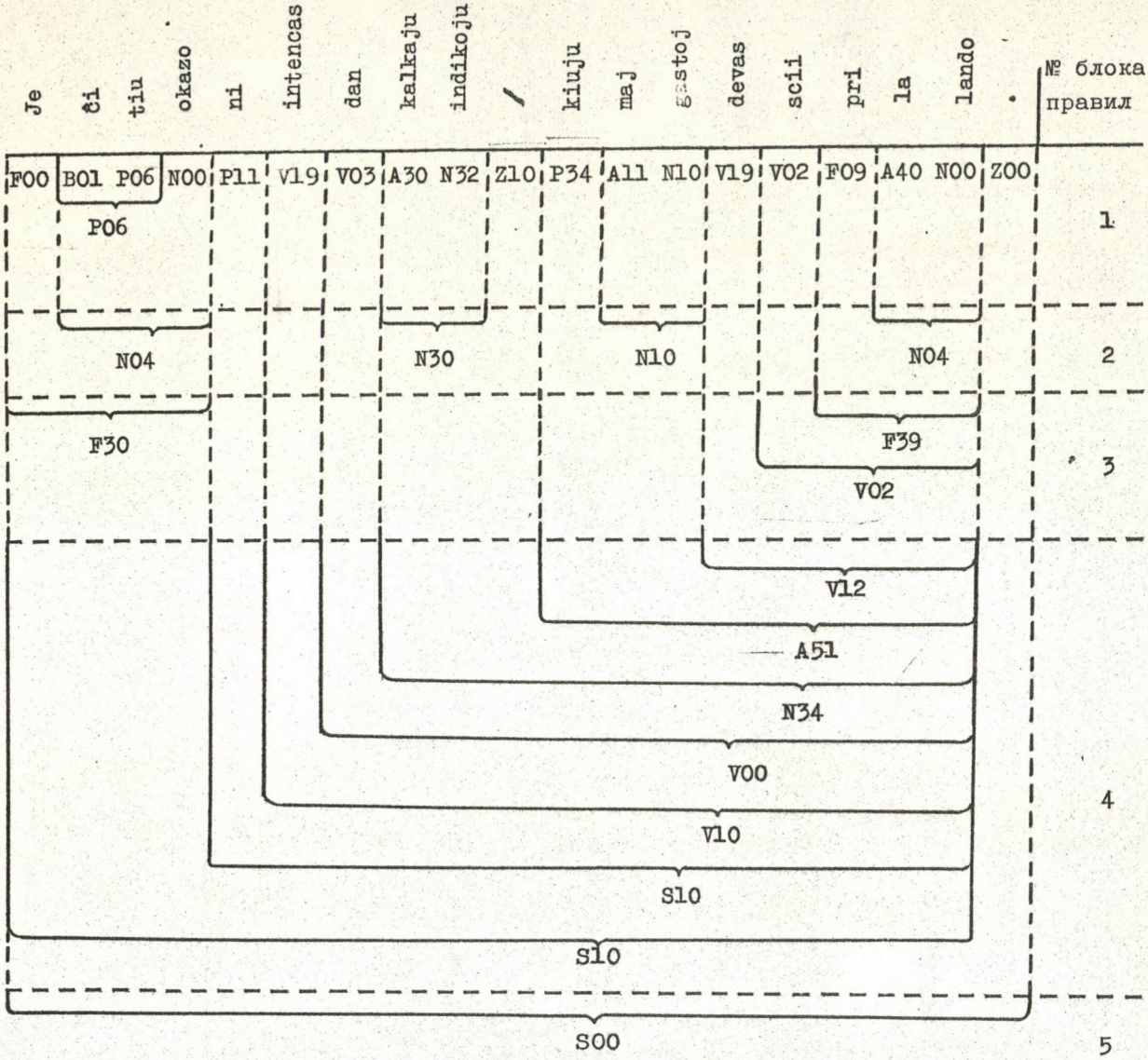


Предложение № 2: La /артикль /51-а /'51-й'/ Universala /'Всемирный'/ Kongreso /'Конгресс'/ okazos /'состоится'/ Budapeŝte /'в Будапеште' наречие/.

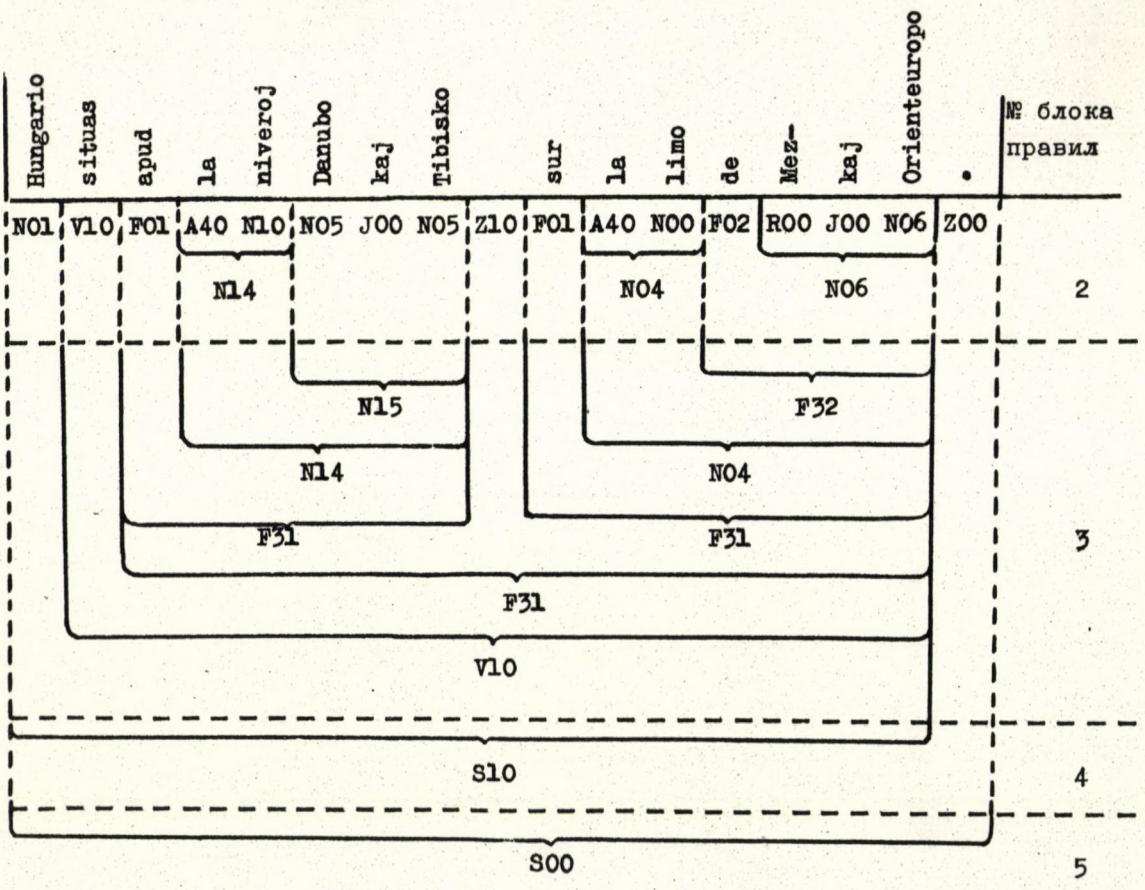


Предложение № 3: Je /предлог неопределенного значения/ ĉi /частица со значением близости/ tiu /'тот'/ okazo /'случай'/ ni /'мы'/ intencas /настоящее время глагола со значением /'намереваться'/ doni /'дать'/ kalkajn /винительный падеж множественного числа прилагательного со значением /'несколько'/ indikojn /'данные', винительный падеж/ kiujn /'которые', вин.пад./ niaj /'наши'/ gasto /'гости'/ devas /настоящее время глагола долженствования/ scii /'знать'/ pri /'о'/ la /артикль/ lando /'страна'/.

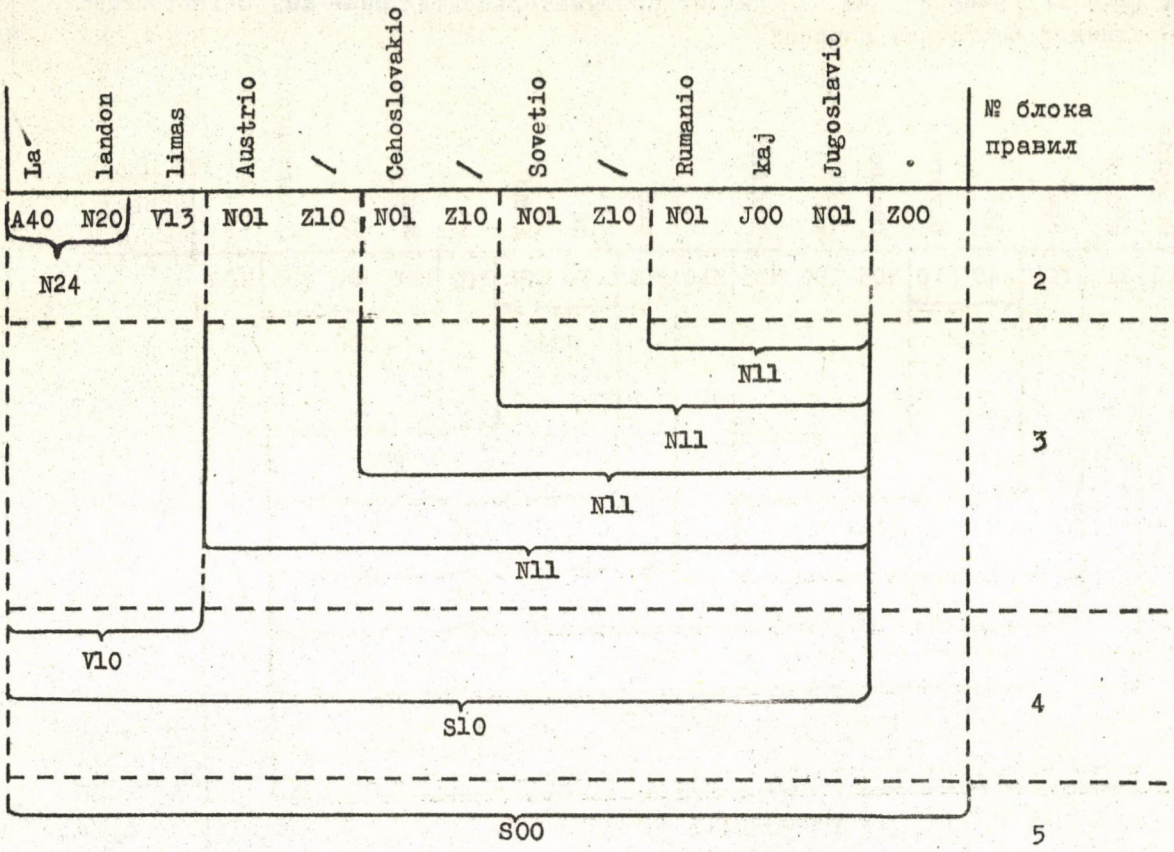
См. дальше на стр. 18.


Je
ĉi
tiu
okazo
ni
intencas
dan
kalkaju
indikoju
/
kiuju
maj
gastoj
devas
scii
pri
la
lando
.
№ блока правил
F00 B01 P06 N00 P11 V19 V03 A30 N32 Z10 P34 A11 N10 V19 V02 F09 A40 N00 Z00
P06
1
N04
N30
N10
N04
2
F30
F39
V02
3
V12
A51
N34
V00
V10
S10
4
S10
S00
5

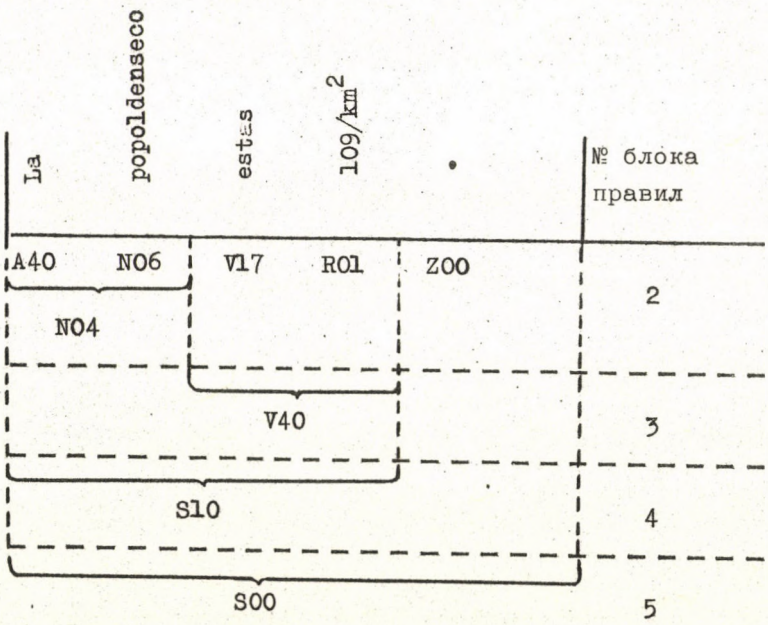
Предложение № 4: Hungario situas /наст.время глагола situi 'быть расположено'/ apud /'возле'/ la riveroj /общий падеж множественного числа от rivero 'река'/ Danubo /'Дунай'/ kaj /'и'/ Tibisko /'Тиса'/ sur /'на'/ la limo /'граница'/ de /предлог притяжательности/ Mez- kaj Orienteuropo /Средняя и Восточная Европа'/



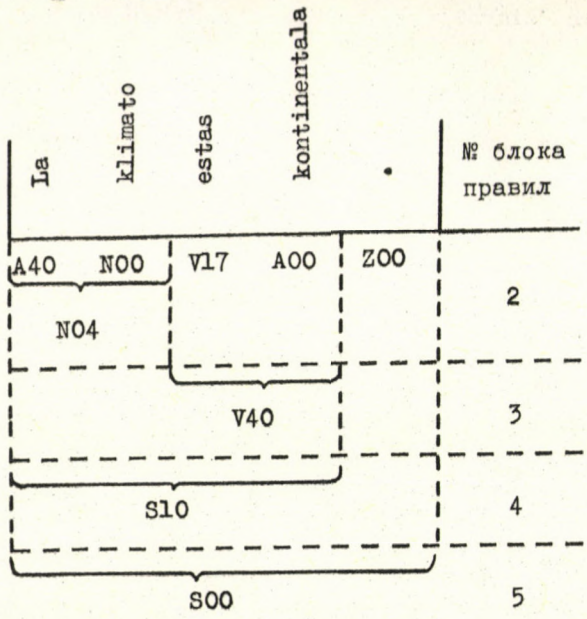
Предложение № 5: La landon /вин.пад.ед.числа/ limas наст. время переходного глагола limi 'граничить'/ Austrio, Cehoslovakio, Sovetio, Rumanio, kaj Jugoslavio.



Предложение № 6: La popoldenseco /'плотность населения'/ estas /наст. время глагола-связки/ 109/km^2.

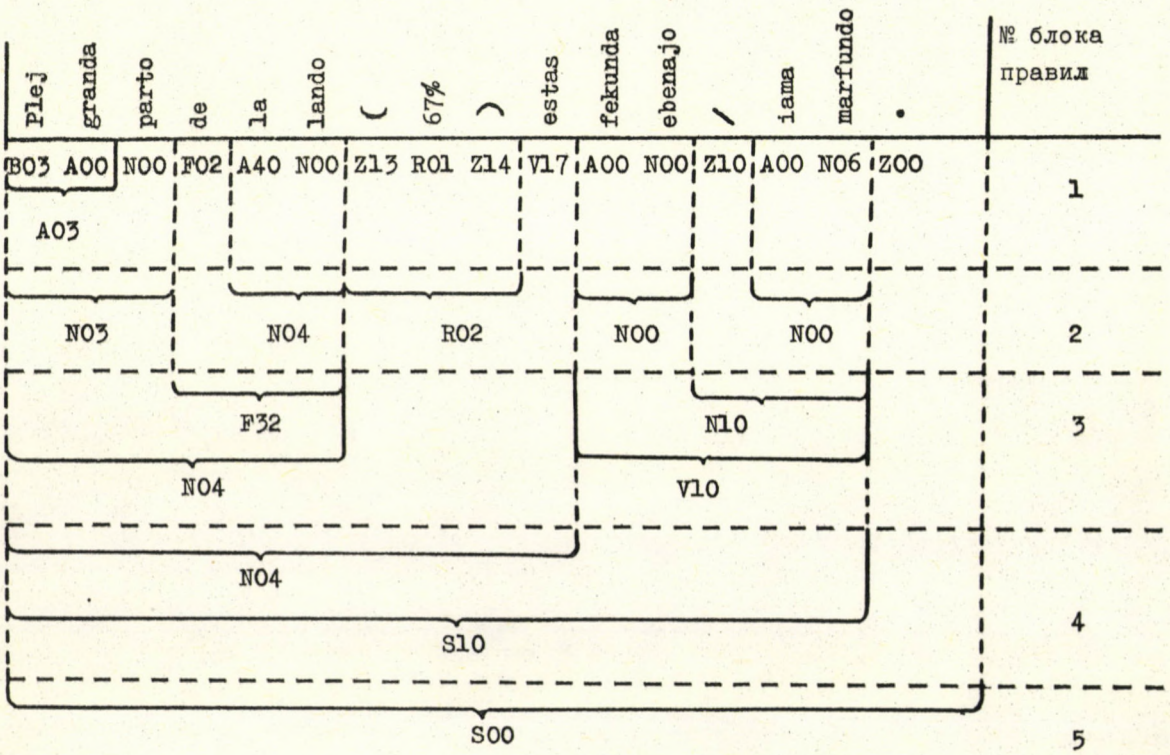


Предложение № 7: La klimato estas kontinentala /общий падеж ед.ч. прилагательного/.

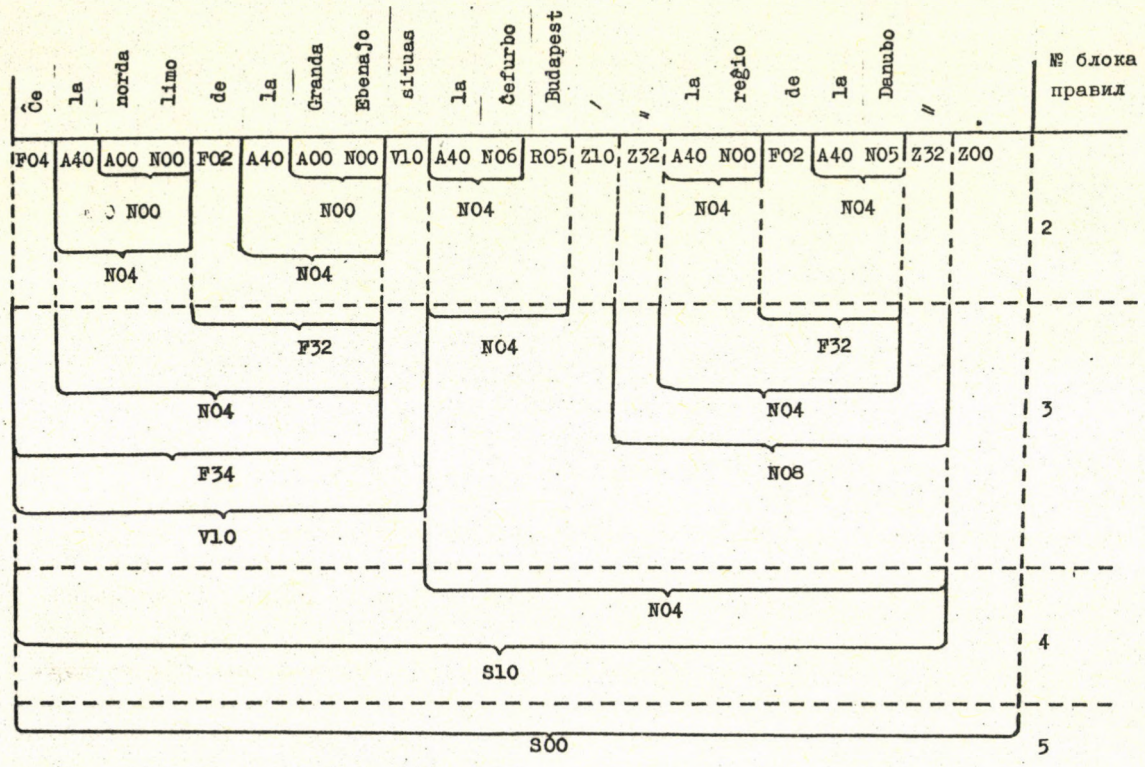


Анализ предложения № 8 похож на анализ двух предыдущих.

Предложение № 9: Plej /частица, указывающая превосходную степень/ granda /'большой', общий пад.ед.ч./ parto /'часть'/ de la lando /67%/ estas fekunda /'плодородный', общий падеж ед.ч./ ebenaĵo /'Низменность', общий падеж ед.ч./, iama /'бывший когда-то', прилагательное в форме общего пад. ед.ч./ marfundo /'дно моря'/.



Предложение № 10: Ce /'у', 'при'/ la norda /'северный'/ limo de la Granda Ebenaĵo situas la ĉefurbo /'столица'/ Budapest, "la reĝino" /'королева'/ de la Danubo".



ВОПРОСЫ СИНТАКСИЧЕСКОГО АНАЛИЗА ДЛЯ ФОРМАЛЬНЫХ ЯЗЫКОВ

Б. Домелки

О Г Л А В Л Е Н И Е

В В Е Д Е Н И Е

1. Формальные языки играют важную роль в автоматизации программирования, а также в решении задач машинного перевода и в других областях прикладной лингвистики. При этом обычно речь идет о языках, определенных с помощью некоторой совокупности формальных правил, выделяющих некоторое множество последовательностей допущенных символов. Такое множество считается языком, определенным данными формальными правилами, а элементы этого же множества называются элементами или предложениями языка.

Задача синтаксического анализа для таких языков состоит в том, чтобы установить о любой последовательности символов, принадлежит ли она к данному языку или нет, т.е. является ли она "правильным" с точки зрения данного языка. Кроме этого, в случае положительного ответа обычно требуется и определение структуры правильной последовательности символов по заданным формальным правилам, т.е. определение последовательности тех формальных правил, применение которых ведет к порождению данного предложения.

В ряде практических случаев формальные правила имеют вид бесконтекстной /контекстно-свободной, context-free/ грамматики. Такая грамматика определяется /см. Хомский [1]/ алфавитом символов A, алфавитом так называемых терминальных символов $A_T \subseteq A$, начальным символом $S \in A \setminus A_T$, и совокупностью правил F. Правила имеют форму упорядоченной пары $(\alpha_K \longrightarrow \beta_K)$, где $\alpha_K \in A \setminus A_T$, а β_K - произвольная непустая последовательность символов алфавита A /т.е. α_K - нетерминальный символ, а β_K слово в алфавите A/.

Элементами языка называются терминальные последовательности /т.е. слова в алфавите A_T/, выводимые из начального символа S с помощью применения конечного числа подстановок, т.е. замещением символов α_K соответствующими словами β_K.

2. Разработан целый ряд различных алгоритмов решения задачи синтаксического анализа для языков, определенных подобными формальными грамматиками. Большинство этих алгоритмов работает аналогично к процессу определения элементов языка: исходя из начального символа S выполняются все возможные подстановки до тех пор, пока не получается искомая последовательность или не исчерпываются все возможности.

Такой подход к анализу называется методом "сверху вниз". Поскольку на каждом шагу может быть применена больше чем одна подстановка, таким образом получается разветвливающийся процесс, реализация которого может потребовать значительного количества времени. Различные варианты алгоритмов такого типа предлагают различные методы организации разветвливания и применяют различные приемы для того, чтобы уменьшить количество безрезультатных проб применения подстановок.

Примеры таких алгоритмов описаны в работах Брукера и Морриса [2] и также Е.А. Жоголева [3]. Они, исходя из начального символа, выполняют одну из возможных подстановок и рассматривают первый /самый левый/ символ получен-

ного слова. Если этот символ является терминальным и совпадает с соответствующим символом анализируемой последовательности, то рассматривается следующий символ обеих последовательностей и т.д. Если получается нетерминальный символ; то выполняется одна из соответствующих подстановок и получается новое слово, на котором весь процесс повторяется. Наконец, если рассматриваемый терминальный символ не совпадает с соответствующим символом анализируемой последовательности, то продолжение данной "ветви" процесса уже не может вести к анализируемой последовательности. В таких случаях придется выбирать другую возможную подстановку вместо последней выполненной, на один уровень выше.

Организация разветвляющегося процесса осуществляется в [2] с помощью рекурсивных обращений к подпрограмме анализа, а в [3]с использованием специального "магазина" (stack) , названного "полем состояний", для запоминания выполненных подстановок. Хотя обе эти алгоритмы имеют ограниченную область действий, в случае практически интересных языков обычно нетрудно преобразовать грамматику в форму, на которую алгоритмы такого типа уже применимы.

В основе метода предсказуемостного анализа см. Куно-Эттингер [4] лежат подобные же принципы: здесь анализ работает тоже в направлении "сверху вниз". Характерная черта этого метода состоит в том, что на каждом шагу выделяются некоторые предсказывания о возможных продолжениях анализируемой последовательности. Эти предсказывания определяются с одной стороны "искомым понятием", т.е. нетерминальным символом, на место которого совершается подстановка, и с другой стороны самым левым символом анализируемой последовательности. В дальнейшем ходе процесса некоторые из этих предсказываний подтверждаются тем, что в конце концов получается анализируемая последовательность.

Универсальность этого метода в области бесконтекстных языков доказывается теоремой Грейбаха /см. [5]/, утверждающей, что к любой бесконтекстной грамматике можно построить эквивалентную грамматику в стандартной форме, где первый символ каждого слова $\beta_к$ является терминальным. Таким образом выполнение каждой подстановки /т.е. применение каждого нового предсказывания/ выделяет новый терминальный символ в начале последовательности, который непосредственно сравнивается с первым символом анализируемой последовательности, так что сразу решается вопрос о применимости данного предсказывания. В результате этого количество действительно выполненных подстановок не превышает длину анализируемой последовательности /не считая, конечно, те, которые соответствуют "опровергнутым" предсказываниям/.

Другой класс алгоритмов основан на работе Айронса [6]. Они работают в направлении "снизу вверх" и исходят из анализируемой последовательности. Рассматривается очередный символ этой последовательности и ищутся подстановки, правая сторона которых начинается этим символом. Для применения такой подстановки /конечно, в обратном направлении/ сначала проверяется, что дальнейшие символы последовательности тоже соответствуют ли очередным символам правой стороны одной из этих подстановок. Если при этом найдется нетерминальный символ на одной из правых сторон, то прежде чем идти дальше, следует провести анализ соответствующей подпоследовательности относительно этого символа. Следовательно, при выполнении процесса анализа по этим методам выделяются различные цели анализа, рекурсивно вызывающие друг друга.

Работа алгоритма значительно ускоряется применением булевой матрицы выводимостей, содержащей единицу в тех и только тех местах, для которых сим-

вол, соответствующий строке, может быть первым символом последовательности, выводимой из символа, соответствующего столбцу. С помощью такой матрицы всегда можно проверить, соответствуют ли очередные символы последовательности актуальной цели анализа и тем самым уменьшить количество безрезультатных проб.

3. Сравнительная оценка некоторых известных методов синтаксического анализа для бесконтекстных языков дается в работе Гриффитса и Пэтрика [7] Экспериментальные данные, собранные в этой работе, доказывают в большинстве случаев преимущество методов типа "снизу вверх" по сравнению методов типа "сверх вниз". Однако, последние все-таки пользуются широкой распространенностью, что объясняется прежде всего тем,что они лучше используют специальные свойства бесконтекстных языков, а именно тот факт, что в случае этих языков возможность применения подстановки проверяется очень просто, так как любой нетерминальный символ $\alpha_к$ может быть заменен на соответствующее слово $\beta_к$. В случае применения методов "снизу вверх" для этого требуется проверить вхождение слова $\beta_к$ и вместо него подставить соответствующий символ $\alpha_к$. Сравнительно большая трудоемкость этого процесса, т.е. проверки вхождений одного из нескольких заданных слов в последовательности символов, является важнейшим препятствием широкого применения методов такого типа.

В главе IV настоящей работы /а также в [13] и [14]/ предложен быстродействующий алгоритм для поиска вхождений слов в последовательности символов. Существование такого алгоритма обеспечивает практическую применимость методов, основанных на принципе "снизу вверх".

Этот принцип в самой чистой форме реализуется в алгоритмах типа "непосредственной подстановки" /direct substitution/ упомянутых в работе [7], как предложенные Абботом и Грейбахом. Здесь подстановки - взятые в обратном направлении - выполняются совершенно независимо от всякой цели анализа, в любом случае, если в анализируемой последовательности найдется вхождение правой стороны некоторой подстановки, и результат рассматривается в качестве анализируемой последовательности. Этот процесс продолжается пока не получается искомый /начальный/ символ, или слово, к которому уже ни одна подстановка не применима.

Легко видеть, что в случае применения этого метода свойство бесконтекстности языка никакого значения не имеет, так как это свойство требует, что левая сторона подстановки, т.е. слово, которое в случае выполнения подстановки заменяет вхождение правой стороны, состоит из одного нетерминального символа. А если уже установлена возможность применения подстановки, то в ее актуальном выполнении не очень существенно, сколько символов придется подставлять.

Таким образом в случае применения метода непосредственной подстановки /с помощью вышеупомянутого быстродействующего алгоритма главы IV/, можно рассматривать не только бесконтекстные языки, а любые системы подстановок вида $(\alpha_к \longrightarrow \beta_к)$, где $\alpha_к$ и $\beta_к$ - произвольные непустые слова в некотором алфавите A.

Этот факт используется для ускорения работы алгоритма в случае бесконтекстных языков тоже: с помощью некоторых преобразований подстановок /об этом речь идет в главе V/, можно значительно уменьшить количество безрезультатных проб /как это в случае методов типа Айронса делается с помощью упомянутой матрицы выводимостей/, но в результате этих преобразования, как правило, теряется свойство бесконтекстности системы.

4. В главе I дается общее определение понятия системы формальных подстановок и связанное с этим определение выводимости слов в таких системах /слово η называется выводимым из слова ξ, если из этого слова, с конечным числом применений подстановок системы, получается слово η/. Решением задачи анализа считается алгоритм, разрешающий о любых двух словах ξ и η, выводимо ли слово η из слова ξ или нет. В этой же главе устанавливается связь с определениями разного типа языков математической лингвистики.

Общая задача анализа не всегда является разрешимой. В главе II вводится понятие ограниченной системы, где длина слов, выводимых из некоторого слова ξ, ограничивается некоторой функцией от этого слова. Если эта функция является в некотором смысле вычислимой, то и задача анализа для соответствующей ограниченной системы будет разрешимой /в таком же смысле/. Далее, если дуальной к системе $(\alpha_к \longrightarrow \beta_к)$ называется система $(\beta_к \longrightarrow \alpha_к)$, то очевидно, что всякая разрешимость системы равносильна разрешимости дуальной системы. Теоремы 2 и 3 дают достаточные условия ограниченности систем; с их помощью можно показать разрешимость задачи анализа для некоторых известных грамматик, в том числе и бесконтекстных и контексно зависящих

являющихся дуальными к ограниченным системам.

К проблемам разрешимости относится и один из результатов главы V. Здесь речь идет о том, что каждая система формальных подстановок определяет некоторый алгоритм преобразования слов в следующем смысле: применяются подстановки к исходному слову ξ, а также к словам, полученным в результате применений подстановок таким образом, что - подобно нормальным алгорифмам А.А. Маркова /см. [8]/ - всегда применяется "первая" в некотором смысле возможность из всех возможностей применения подстановок. Но в то время как при нормальных алгорифмах всегда выбирается первая по упорядоченности подстановка и самое левое ее вхождение, то здесь ищется самое левое вхождение любой подстановки и только в случае нескольких подстановок, заканчивающихся на одном и том же символе, выбирается из них первая по упорядоченности подстановка.

Теорема 7 утверждает равносильность алгоритмов такого типа к нормальным алгорифмам и так всем видам эффективно выполнимых алгоритмов.

Для упрощения решения задачи анализа служат две теоремы типа редукции, утверждающие, что, если слово η выводимо из слова ξ, то всегда существует и вывод специальной формы между этими двумя словами. В случае теоремы 1 такими специальными выводами являются бесповторные выводы, содержащие только различные слова, а в случае теоремы 4 последовательные выводы, где ни одна примененная подстановка не должна заканчиваться полностью левее от начала слова, подставленного на предыдущем шагу.

Алгоритм анализа, изложенный в главе III, работает путем просмотра дерева таких бесповторных и последовательных выводов. Узлами этого дерева являются слова, выводимые из некоторого исходного слова ξ, а ветвями - различные возможности применения подстановок к этим словам. Имеются в виду только те возможности, которые не нарушают вышеуказанные свойства получаемых выводов. Упорядочение ветвей, связанных с одним и тем же узлом, соответствует принципу, упомянутому в связи с понятием алгоритма теоремы 7.

Просмотр дерева осуществляется таким образом, что из каждого узла просмотр продолжается по первой /самой левой/ ветви, а если продолжение некоторого пути оказывается бесперспективным /т.е. искомое слово η уже не может быть выведено по этому пути; такой бесперспективный путь будем называть

"тупик"-ом/, то выбирается следующая ветвь из последнего /разветвливающегося/ узла.

Тот факт, что рассматриваются только последовательные выводы и вышеуказанный метод упорядочения ветвь обеспечивают, что поиск новой применяемой подстановки всегда начинается "правее" от начала последней выполненной подстановки. Таким образом ограничивается количество повторно просматриваемых символов.

Это соответствует вышеупомянутому методу поиска вхождений слов, изложенному в теореме 6 главы IV. В основе методы лежат операции над булевыми векторами, исходя из того факта, что электронные цифровые вычислительные машины выполняют логические операции над целыми машинными словами, т.е. одновременно над совокупностями независимых логических значений.

Левые стороны подстановок системы, т.е. слова $\alpha_к$, вхождения которых ищутся, изображаются в форме булевой матрицы порядка $(n \times \rho)$, где n - количество элементов алфавита системы, а ρ - сумма длин слов $\alpha_к$. С помощью простой рекурсивной формулы, для каждого символа анализируемой последовательности вычисляется значение некоторого булевого вектора Q, длины ρ, каждый разряд которого соответствует некоторому символу одного из слов $\alpha_к$. Единица в этом разряде означает, что в анализируемой последовательности найдено начало слова $\alpha_к$, вплоть до этого символа. Таким образом, единица в разряде, соответствующем концу некоторого слова $\alpha_к$, означает вхождение этого слова в анализируемую последовательность. Ситуация $Q = \underline{0}$ соответствует тупику, что можно показать с использованием теоремы 5 главы III.

В ходе работы алгоритма анализа значения векторов Q запоминаются, чтобы обеспечивать возможность к восстановлению старой ситуации в случае продолжения процесса после выполнений подстановок или для выхода из тупика.

Следует подчеркнуть, что практическая применимость предлагаемого метода анализа типа "непосредственной подстановки" для общих систем формальных подстановок на современных ЭЦВМ обусловливается именно существованием такого быстродействующего метода для поиска возможностей применений подстановок, с одновременным запоминанием старых ситуаций и своевременным обнаруживанием тупиков.

"Узким местом" предлагаемого алгоритма - с точки зрения требуемого машинного времени - несомненно является возможное большое количество тупиков. С целью устранения этого недостатка вводятся в главе V понятие прямого вывода, выполняющего всегда первую возможную подстановку, и понятие финальной подстановки, применение которой всегда считается окончательным в том смысле, что если применение такой подстановки вел бы в тупик, то и применение вместо нее любой другой подстановки тоже должно вести в тупик. Легко видеть, что если все подстановки системы являются финальными, то все выводы должны быть прямыми, т.е. не должны допускать тупиков.

Теоремы 8 и 9 дают достаточное /и в некотором смысле необходимое/ условие финальности подстановок, и указаны некоторые приемы "финализации" подстановок, путем добавления к обеим сторонам подстановки некоторого контекста, обеспечивающего выполнение подстановки только в тех случаях, если это не ведет в тупик. Пример такой финализации системы детально разработан в конце главы V.

В главе VI рассматривается возможность применения предлагаемых методов в трансляторе с языка АЛГОЛ-60. Синтаксис этого языка переработан в систему формальных подстановок, причем все подстановки этой системы являются фи-

нальными, и так в ходе анализа тупиков не могут быть. При анализе имеются в виду различные типы и виды идентификаторов, что осуществляется предварительным просмотром для обработки описаний. Этот первый просмотр выполнен тоже с помощью метода синтаксического анализа.

Предлагается такая организация транслятора, где актуальный перевод производится сематическими подпрограммами, вызываемыми программой синтаксического анализа при нахождении соответствующих синтаксических понятий.

Настоящая работа является диссертацией автора на соискание ученой степени кандидата физико-математических наук, представленной на механико-математическом факультете Московского Государственного Университета в мае 1966 года. Пользуясь случаем, автор выражает глубокую благодарность своему научному руководителю, профессору М.Р. Шура-Бура за постоянную помощь в разработке диссертации, опонентам С.С. Лаврову и В.В. Мартынюку за их ценные замечания и Е.А. Жоголеву за некоторые предложения в связи с главой VI.

Программы некоторых частей алгоритма анализа были отлажены на машине М-20 Вычислительного Центра МГУ. На основе предложенного алгоритма в настоящее время разрабатывается синтаксически-управляемый транслятор с языка АЛГОЛ-60 для машины МИНСК-2 в Центральном Статистическом Управлении ВНР. В этой работе кроме автора принимают участие Ж. Рэмэтэи и Т. Бакош. Кроме того, в Вычислительном Центре Венгерской Академии Наук Д. Варга применяет предложенный метод в экспериментальной программе для синтаксического анализа предложений русского языка на машине УРАЛ-2. Автор благодарен всем упомянутым за их помощь и замечания в связи с реализацией алгоритма.

I. ОСНОВНЫЕ ПОНЯТИЯ О СИСТЕМАХ ФОРМАЛЬНЫХ ПОДСТАНОВОК

1. Пусть $A = \{\omega_1, \omega_2, \ldots, \omega_n\}$ - конечный алфавит символов. Множество всех слов, построенных из элементов алфавита A, обозначается через $S(A)$

/1.1/ $\xi = \omega_{i_1}\omega_{i_2}\ldots\omega_{i_q} \in S(A) \quad (1 \le i_j \le n)$.

Длина слова ξ обозначается через $|\xi| = q$.
Пустое слово обозначается через Λ, причем $|\Lambda| = 0$.
Символ ω_{i_j}, т.е. j -тый символ слова ξ, иногда будем называть символом порядка j слова ξ.

Слова в алфавите A обозначаются греческими, множества таких слов прописными латинскими и натуральные числа - строчными латинскими буквами. Предполагается, что все символы, использованные в "мета" - смысле /например, скобки, запятые, $\longrightarrow$, = и т.д./ не принадлежат к алфавиту A.

2. Определяется понятие <u>сегмента</u> слова ξ :

/1.2/ $$[\xi]_p^r = \begin{cases} \omega_{i_{p+1}}\omega_{i_{p+2}}\ldots\omega_{i_r}, & \text{если } 0 \le p < r \le |\xi| \\ \Lambda & \text{иначе,} \end{cases}$$

и введутся обозначения $[\xi]^r = [\xi]_0^r$ и $[\xi]_p = [\xi]_p^{|\xi|}$.

Из определения непосредственно следует, что для $0 \le p < r < |\xi|$

/1.3/ $\xi = [\xi]^p [\xi]_p^r [\xi]_r$.

3. Упорядоченная пара различных слов $\alpha \in S(A)$ и $\beta \in S(A)$ при $\alpha \neq \beta$ называется <u>формальной подстановкой в алфавите A</u> и обозначается через $(\alpha \to \beta)$.
Если не сказано противоположное, предполагается, что $\alpha \neq \Lambda$ и $\beta \neq \Lambda$
формальная подстановка

/1.4/ $F = (\alpha \to \beta)$

называется <u>применимой</u> к символу порядка S слова ξ, если

/1,5/ $[\xi]_{s-|\alpha|}^{s} = \alpha$

В этом случае определяется <u>оператор применения подстановки</u> F на символ порядка S слова ξ следующим образом:

/1.6/ $F^s(\xi) = [\xi]^{s-|\alpha|}\beta[\xi]_s$.

Если /1.5/ не имеет места, то результат применения оператора $F^s(\xi)$ считается неопределенным.

Некоторые свойства этого оператора явно следуют из определения /см. рис. 1/

/1.7/ если $r \le s-|\alpha|$ то $[F^s(\xi)]^r = [\xi]^r$

/1.8/ если $\rho \ge s$ то $[F^s(\xi)]_{\rho-|\alpha|+|\beta|} = [\xi]_\rho$

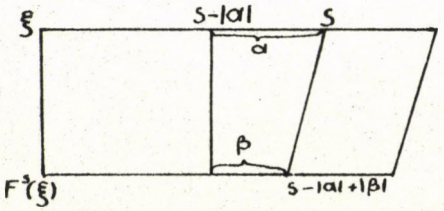


Рис. 1

Далее, пусть формальная подстановка $(\beta \to \alpha)$ называется дуальной к подстановке $F=(\alpha \to \beta)$ и обозначается через $\overline{F}$. Легко видеть, что

/1.9/ если $\eta = F^s(\xi)$, то $\xi = \overline{F}^{s\cdot|\alpha|+|\beta|}(\eta)$.

4. Конечный набор формальных подстановок над алфавитом A, $F=\{F_1, F_2, \ldots\ldots, F_t\}$ называется системой формальных подстановок над алфавитом A /в дальнейшем просто системой/. Определяется понятие выводимости относительно системы F:

Если к словам ξ и η, при $\xi, \eta \in S(A)$, найдутся числа s и k такие, что $\eta = F_k^s(\xi)$, то слово η называется непосредственно выводимым из слова ξ , и этот факт обозначается через

/1.10/ $F: \xi \vdash \eta$.

Далее, если существуют слова $\xi_1, \xi_2, \ldots\ldots, \xi_{t-1}$, такие, что

/1.11/ $F: \xi = \xi_0 \vdash \xi_1 \vdash \ldots\ldots \vdash \xi_{t-1} \vdash \xi_t = \eta$,

то слово η называется выводимым из слова ξ ; этот факт обозначается через

/1.12/ $F: \xi \models \eta$,

и /1.11/ называется выводом слова η . Легко видеть, что если $F: \xi \models \eta$, то $\overline{F}: \eta \models \xi$. Этот факт в дальнейшем будем называть принципом дуальности.

Множество всех слов η , для которых имеет место /1.12/ обозначается через $\mathcal{M}_F(\xi)$.

5. Понятие выводимости в системе формальных подстановок F над алфавитом A показывает большое сходство с фразево-структурными языками математической лингвистики /см. Хомский [1]/. В самом деле, пусть $\Gamma = (A, A_T, S, F)$ язык, который определен следующим образом: конечные множества A и A_T называются алфавитами символов и терминальных символов соответственно, $S \in A \setminus A_T$ есть начальный символ и F - множество упорядоченных пар слов α_k и β_k в алфавите A.

Слова в алфавите A_T, получаемые из начального символа S с помощью конечного числа подстановок слов β_k вместо вхождений соответствующих слов α_k , называются элементами или предложениями языка.

Легко видеть, что между множеством L_Γ всех предложений языка и системой формальных подстановок F над алфавитом A имеет место следующее соотношение :

/1.13/ $L_\Gamma = \mathcal{M}_F(S) \cap S(A_T)$.

Таким образом различными частными случаями фразево-структурных языков соответствуют некоторые системы формальных подстановок. Важнейшими частными случаями являются контексно-свободные языки, определенные соотношением

/1.14/ $\alpha_k \in A \setminus A_T$, $\beta_k \neq \Lambda$

и контексно-зависящие языки, где

/1.15/ $\alpha_k = \gamma_k \omega_{i_k} \tau_k$, $\beta_k = \gamma_k \delta_k \tau_k$

при

$\omega_{i_k} \in A \setminus A_T$, $\delta_k \neq \Lambda$, γ_k , $\tau_k \in S(A)$

II. ЗАДАЧА АНАЛИЗА В СИСТЕМАХ ФОРМАЛЬНЫХ ПОДСТАНОВОК

1. Пусть даны система F над алфавитом A и слова ξ и η в этом же алфавите. Задача анализа заключается в решении вопроса, является ли слово η выводимым из слова ξ, т.е. имеет ли место

/2.1/ $\eta \in \mathcal{M}_F(\xi)$

В случае самых общих систем F эта задача является алгоритмически неразрешимой: можно показать эквивалентность этой задачи с известной проблемой тождества слов в ассоциативных системах /см. Марков [8]/.

Общая неразрешимость задачи не исключает возможность существования алгоритмов для решения некоторых частных случаев. Так например, если можно показать, что количество выводов, которые могут дать в результате некоторое слово η, является конечным, то перебор всех таких выводов и будет искомым алгоритмом.

Для сокращения количества рассматриваемых выводов вводится понятие бесповторного вывода.

2. Вывод

/2.2/ $F: \xi = \xi_0 \vdash \xi_1 \vdash \ldots \vdash \xi_{t-1} \vdash \xi_t = \eta$

называется <u>бесповторным</u>, если все его члены различны, т.е. $\xi_i \neq \xi_j$ для $0 \le i < j \le t$

<u>Теорема 1</u>

Если $\eta \in \mathcal{M}_F(\xi)$, то существует и бесповторный вывод слова η из слова ξ

<u>Доказательство:</u>

Предложим, что в выводе /2.2/ найдутся одинаковые слова $\xi_i = \xi_j$, при $0 \le i < j \le t$. В этом случае вычеркиванием слов $\xi_{i+1}, \xi_{i+2}, \ldots \xi_j$ получается вывод

/2.3/ $F: \xi = \xi_0 \vdash \xi_1 \vdash \ldots \vdash \xi_i = \xi_j \vdash \xi_{j+1} \vdash \ldots \vdash \xi_t = \eta$

имеющий длину $t-(i-j) < t$. Таким образом конечным повторением этого процесса получается бесповторный вывод слова η из слова ξ

Таким образом в дальнейшем без ограничения общности можем рассматривать только бесповторные выводы.

3. Система F называется <u>ограниченной</u>, если существует функция $b_F(\xi)$, сопоставляющая каждому слову $\xi \in S(A)$ некоторое натуральное число, такое, что для любого слова $\eta \in \mathcal{M}_F(\xi)$ имеет место

/2.4/ $|\eta| \le b_F(\xi)$

Для ограниченных систем задача анализа действительно решается конечным перебором: для данного слова ξ рассматриваются только те выводы, которые содержат слова с длиной не больше чем $b_F(\xi)$. Поскольку количество таких слов, и значит количество элементов множества не может быть больше чем $(n^{b_F(\xi)+1} - 1)/(n-1)$,

где n количество элементов алфавита A, в силу бесповторности выводов и длина вывода будет ограниченной.

Легко видеть, далее, что в случае /примитивно/ рекурсивной функции $b_F(\xi)$ и сам этот процесс конечного перебора будет /примитивно - / рекурсивным, т.е. характеристическая функция

/2.5/ $$r_F(\xi, \eta) = \begin{cases} 0, & \text{если } \eta \in M_F(\xi) \\ 1, & \text{иначе} \end{cases}$$

тоже оказывается /примитивно -/ рекурсивной. /О рекурсивности функций, определенных над словами, см. Петер [9]. Примитивно-рекурсивность задачи анализа для контексно-свободных языков доказана в работе Петер [10]/

В силу принципа дуальности

/2.6/ $r_F(\xi, \eta) = r_F(\eta, \xi)$,

таким образом ограниченность системы $\overline{F}$ обеспечит возможность решения конечным перебором /или рекурсивность/ задачи анализа для системы F тоже.

4. Ниже приводятся некоторые достаточные условия для ограниченности системы.

Пусть целочисленная функция $\rho(\xi) \geq 0$, где $\rho(\xi) = 0$ имеет место только для $\xi = \Lambda$, называется <u>нормой</u>. Норма $P(\xi)$ называется <u>аддитивной</u>, если для любых двух слов ξ и η

/2.6/ $\rho(\xi\eta) = \rho(\xi) + \rho(\eta)$

<u>Теорема 2</u>

Если подстановки системы F для некоторой аддитивной нормы P удовлетворяют условию

/2.7/ $\rho(\alpha_k) \geq \rho(\beta_k)$ $\qquad (1 \leq k \leq f)$

то система является ограниченной при $b_F(\xi) = \rho(\xi)$.

<u>Доказательство:</u>

Из аддитивности нормы и из /2.6/ следует, что применение подстановки не увеличивает норму слова. Таким образом

/2.8/ $\rho(\eta) \leq \rho(\xi)$ $\qquad (\eta \in M_F(\xi))$

С другой стороны, для каждого символа ω имеет место $\rho(\omega) \geq 1$ и, таким образом, если $\eta = \omega_{i_1} \omega_{i_2} \ldots \omega_{i_{|\eta|}}$, то

/2.9/ $\rho(\eta) = \sum_{j=1}^{|\eta|} \rho(\omega_{i_j}) \geq |\eta|$

Из /2.8/ и /2.9/ получается /2.4/, для $b_F(\xi) = \rho(\xi)$.

Самым простым примером аддитивной нормы может служить <u>длина</u> слова. Таким образом теорема 2 обеспечивает ограниченность систем, дуальных к контексно-свободным /бесконтекстным/ и к контексно-зависящим языкам, откуда согласно сказанному в п.3 и следует возможность конечного и примитивно-рекурсивного решения задачи анализа для таких языков.

5. В том, что теорема 2 дает только достаточное условие ограниченности, можно убедиться с помощью простого примера

/2.10/ $F = \{(a \to bcd), (bd \to a)\}$

Здесь аддитивная норма, удовлетворяющая условию /2.7/ невозможна, так как требовалось бы

$$\rho(a) \geq \rho(b) + \rho(c) + \rho(d) > \rho(b) + \rho(d) \geq \rho(a)$$

С другой стороны, ограниченность системы очевидна.

6. В некоторых таких случаях ограниченность системы получается из другой теоремы:

<u>Теорема 3</u>

Если для системы F существует норма $\rho(\xi)$ такая, что для любых двух слов γ и γ

/2.11/ $\rho(\gamma\alpha_\kappa\gamma) > \rho(\gamma\beta_\kappa\gamma) \qquad (1 \le \kappa \le m)$

то система является ограниченной при

/2.12/ $b_\rho(\xi) = |\xi| + \rho(\xi) \cdot \max(0, \max_\kappa(|\beta_\kappa| - |\alpha_\kappa|))$

Доказательство:

Из /2.11/ следует, что каждый шаг вывода уменьшает норму. Таким образом, длина вывода не может быть больше чем $\rho(\xi)$. С другой стороны каждый шаг может увеличить длину слова не больше чем $\max_\kappa(|\beta_\kappa| - |\alpha_\kappa|)$, если это число является положительным. Таким образом для длины слов, выводимых из ξ, /2.12/ действительно имеет место.

7. Следует отметить, что в условии /2.11/ знак неравенства нельзя заменить на "$\ge$", как это было в случае аддитивной нормы в теореме 2. Действительно для тривиальной /неаддитивной/ нормы $\rho(\xi) \equiv 1$ для любого ξ имеет место

/2.13/ $P(\gamma\alpha_\kappa\gamma) = \rho(\gamma\beta_\kappa\gamma)$

для любой системы.

8. В некоторых случаях теорема 3 действительно оказывается более мощной чем теорема 2. Так, например, в системе примера /2.10/ можно определить следующую норму

$$\rho(a)=3, \quad \rho(b)=\rho(c)=\rho(d)=2, \quad \rho(bcd)=2 ,$$

а в остальных /кроме слова bcd/ норма будет аддитивной. Так как ни одно из слов $\gamma a\gamma, \gamma bcd\gamma, \gamma bd\gamma$ не может содержать новое вхождение "нерегулярного" слова bcd, то из-за аддитивности нормы для остальных слов /2.11/ действительно получается.

9. С другой стороны, примером системы, ограниченность которой показывается теоремой 2, а не теоремой 3, может служить

/2.14/ $F = \{(a \to b), (b \to a)\}$.

С помощью аддитивной нормы $\rho(\xi) = |\xi|$ теорема 2 очевидно применима, применение же теоремы 3 очевидно невозможно.

Таким образом теоремы 2 и 3 дают независимо друг от друга достаточные условия ограниченности системы.

III. АЛГОРИТМ АНАЛИЗА, ОСНОВАННЫЙ НА СИСТЕМАТИЧЕСКИЙ ПЕРЕБОР ВОЗМОЖНОСТЕЙ

1. Теоремы 2 и 3 обеспечивают возможность решения задачи анализа в некоторых случаях /включая случай анализа более важных для практики языков/ с помощью перебора всех возможностей, число которых ограничивается этими теоремами. В этой главе рассматривается вопрос эффективного выполнения этого перебора, путем исключения некоторых возможностей, о которых можно показать, что они являются в некотором смысле "неинтересными" с точки зрения решения задачи.

Для этого, во-первых, определяется специальный класс выводов, о которых - подобно бесповторным выводам - можно доказать, что с их помощью можно получить все выводимые слова.

2. Пусть рассматривается вывод

/3.1/ $F: \xi = \xi_0 \vdash \xi_1 \vdash \ldots \vdash \xi_{t-1} \vdash \xi_t = \eta;$

и числа $s_1, s_2, \ldots, s_t$ $k_1, k_2, \ldots, k_t$ определяются соотношением

/3.2/ $\xi_i = F_{k_i}^{s_i}(\xi_{i-1})$ $(1 \le i \le t)$

Будем называть вывод <u>r - последовательным</u>, если для $0 \le i \le t-1$ имеет место

/3.3/ $$S_{i+1} > r_i = \begin{cases} r, & \text{если } i=0 \\ S_i - |\alpha_{k_i}|, & \text{если } i \ge 1 \end{cases}$$

Для $r = 0$, 0 - последовательные выводы будем называть просто последовательными. Множество всех слов выводимых r - последовательно из слова ξ, обозначается через $U_F^r(\xi)$

Значение последовательных выводов для процесса анализа заключается в том, что после выполнения одной подстановки, следующую подстановку можно искать только "правее" от начала той части слова, которая принимала участие в выполненной подстановке.

Об универсальности последовательных выводов можно доказать следующую теорему:

<u>Теорема 4</u>

Для любой системы F и слова $\xi \in S(A)$ имеет место

/3.4/ $M_F(\xi) = U_F^0(\xi)$

т.е. для каждого выводимого слова существует последовательный вывод.

<u>Доказательство:</u>

Пусть i-тый шаг будет первым непоследовательным шагом некоторого вывода:

/3.5/ $S_{i+1} \le r_i$, но $s_{i+1} > r_j$ для $0 \le j \le i-1$

Из определения оператора применения подстановки получается

/3.6/ $\xi_i = [\xi_{i-1}]^{r_i} \beta_{k_i} [\xi_{i-1}]_{s_i} = [\xi_i]^{r_{i+1}} \alpha_{k_{i+1}} [\xi_i]_{s_{i+1}}$

и так согласно /1.7/ и /3.5/

/3.7/ $[\xi_i]^{r_{i+1}} \alpha_{k_{i+1}} = [\xi_i]^{s_{i+1}} = [\xi_{i-1}]^{s_{i+1}}$,

т.е. подстановка $(\alpha_{k_{i+1}} \to \beta_{k_{i+1}})$ оказывается применимой уже к слову ξ_{i-1}. Если при этом все еще $S_{i+1} \le r_{i-1}$, то таким образом из

/3.8/ $\xi_{i-1} = [\xi_{i-2}]^{r_{i-1}} \beta_{k_{i-1}} [\xi_{i-2}]_{s_{i-1}}$,

и /3.7/ следует

/3.9/ $[\xi_i]^{r_{i+1}}\ \alpha_{k_{i+1}} = [\xi_i]^{s_{i+1}} = [\xi_{i-1}]^{s_{i+1}} = [\xi_{i-2}]^{s_{i+1}}$

Пусть этот процесс продолжается до тех пор, пока не найдется число j такое, что

/3.10/ $s_{i+1} > r_{i-j}$, но $s_{i+1} \le r_l$, $(i-j+1 \le l \le i)$.

/см. рис. 2/. Такое число j существует, так как $r_0 = 0$.

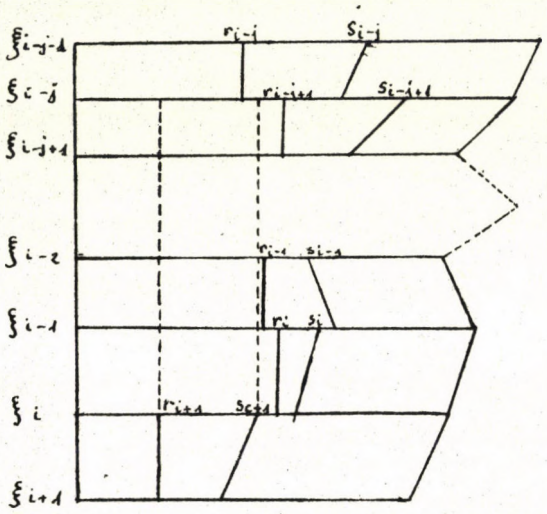


Рис. 2

Таким образом, подобно предыдущим имеет место

/3.11/ $[\xi_i]^{r_{i+1}}\ \alpha_{k_{i+1}} = [\xi_i]^{s_{i+1}} = \ldots = [\xi_{i-j}]^{s_{i+1}}$,

и так подстановка $(\alpha_{k_{i+1}} \to \beta_{k_{i+1}})$ применима к слову ξ_{i-j}, а из-за /3.10/ это применение уже не нарушает последовательности вывода.

Строится вывод $\{\xi'_l\}$, отличающийся от оригинального вывода $\{\xi_l\}$ только для $i-j+1 \le l \le i$.

Пусть, во-первых

/3.12/ $\xi'_{i-j+1} = F^{s_{i+1}}_{k_{i+1}}(\xi_{i-j}) = [\xi_{i-j}]^{r_{i+1}}\ \beta_{k_{i+1}}[\xi_{i-j}]_{s_{i+1}}$

а далее покажем, что все подстановки, примененные в течении оригинального вывода к слову ξ_{i-j}, применимы в таком же порядке к слову ξ'_{i-j+1}

Для этого пусть будет , и

/3.13/
$$s'_{l+1} = \begin{cases} s_{i+1}, & \text{если } l = i-j \\ s_l - d_l, & \text{если } i-j+1 \le l \le i \end{cases}$$

$$k'_{l+1} = \begin{cases} k_{i+1}, & \text{если } l = i-j \\ k_l & \text{если } i-j+1 \le l \le i \end{cases}$$

$$\xi'_{l+1} = F^{s'_{l+1}}_{k'_{l+1}}(\xi'_l) \qquad (i-j \le l \le i)$$

Доказывается, что для $i-j+1 \le l \le i+1$ имеет место

/3.14/ $\xi'_l = [\xi_{i+1}]^{s_{i+1}-d}\ [\xi_{l-1}]_{s_{i+1}}$,

откуда из $s_{i+1} \le r_l$ получается

/3.15/ $[\xi'_\ell]_{r_\ell - d}^{s_\ell - d} = [\xi_{\ell-1}]_{r_\ell}^{s_\ell} = \alpha_{\kappa_\ell} = \alpha_{\kappa'_{\ell+1}}$

т.е. применимость подстановки $(\alpha_{\kappa'_{\ell+1}} \to \beta_{\kappa'_{\ell+1}})$ к слову ξ_ℓ.

Действительно, для $\ell = i-j+1$ из /3.11/ и /3.12/ сразу следует /3.14/. Предположим, что /3.14/ уже доказано для некоторого числа ℓ. В этом случае из /3.13/ и /3.15/

$$\xi'_{\ell+1} = F_{\kappa'_{\ell+1}}^{s'_{\ell+1}}(\xi'_\ell) = [\xi'_\ell]^{r_\ell - d}\ \beta_{\kappa_\ell}[\xi'_\ell]_{s_\ell - d}$$

отсюда из /3.14/ и $s_{i+1} \le r_\ell$ получается

/3.16/ $$\xi_{\ell+1} = [\xi_{i+1}]^{s_{i+1}-d}[\xi_{\ell-1}]_{s_{i+1}}^{r_\ell}\ \beta_{\kappa_\ell}[\xi_{\ell-1}]_{s_\ell} = [\xi_{i+1}]^{s_{i+1}-d}\ [\xi_\ell]_{s_{i+1}}$$

т.е. /3.14/ имеет место и для числа $\ell+1$, и так по индукции для всех чисел $i-j+1 \le \ell \le i+1$.

Таким образом, для $\ell = i+1$ имеет место

/3.17/ $$\xi'_{i+1} = [\xi_{i+1}]^{s_{i+1}-d}\ [\xi_i]_{s_{i+1}} = \xi_{i+1}$$

и так для $\ell \ge i+1$ оба вывода $\{\xi_\ell\}$ и $\{\xi'_\ell\}$ совпадают. Но в выводе ξ'_ℓ все первые i шагов уже будут последовательными, так как $s'_{i-j+1} = s_{i+1} > r_{i-j}$ согласно /3.10/, а далее

/3.18/ $$s'_{i-j+2} = s_{i-j+1} - |\alpha_{\kappa_{i+1}}| + |\beta_{\kappa_{i+1}}| > s_{i-j+1} - |\alpha_{\kappa_{i+1}}| >$$

$$> r_{i-j+1} - |\alpha_{\kappa_{i+1}}| \ge s_{i+1} - |\alpha_{\kappa_{i+1}}| = r_{i+1} = r_{i-j+1}\ .$$

а для $i-j+2 \le \ell \le i$ из последовательности первых $i-1$ шагов оригинального вывода получается

/3.19/ $$s'_{\ell+1} = s_\ell - d > r_{\ell-1} - d = r'_\ell$$

Таким образом первый непоследовательный шаг вывода $\{\xi'_\ell\}$ уже находится после i-того шага и так повторением этого процесса можно исключить все нарушения последовательности.

Замечание

В /3.18/ было использовано условие $\alpha_\kappa \ne \Lambda$ и $\beta_\kappa \ne \Lambda$.

Действительно, в случае $\beta_{\kappa_{i+1}} = \Lambda$ первое, а в случае $\alpha_{\kappa_{i-j+1}} = \Lambda$ второе неравенство превратится в равенство. Имея в виду, что для доказательства последовательности вывода $\{\xi'_\ell\}$ достаточно выполнение хотя бы одного из этих неравенств, теорему 4 можно доказать и при более слабых условиях, как например требование того, что если $\alpha_\ell = \Lambda$, то для всех k, $\beta_\kappa \ne \Lambda$, или наоборот.

3. В силу только что доказанной теоремы если ищется какой-то вывод, то после выполнения подстановки $\xi_i = F_{\kappa_i}^{s_i}(\xi_{i-1})$ конец следующей подстановки можно искать "правее" от начала подстановленного слова, т.е. от порядка r_i символа слова ξ_i. Таких возможностей применения подстановок может быть много. Вводится упорядочение этих возможностей. Пусть наименьшее число u, для которого существует такое число ℓ, что

/3.20/ $$u > r \quad \text{и} \quad [\xi]_{u-|\alpha_\ell|}^{u} = \alpha_\ell$$

обозначается через $u_r^1(\xi)$, а соответствующее число ℓ через $\ell_r^1(\xi)$. Далее, если числа $u_r^j(\xi)$ и $\ell_r^j(\xi)$ уже определены, то наименьшее число u, для которого существует такое число ℓ, что $[\xi]_{u-|\alpha_\ell|}^{u} = \alpha_\ell$ и

/3.21/ $u > u_r^j(\xi)$ или, если $u = u_r^j(\xi)$ то $\ell > \ell_r^j(\xi)$,

обозначается через $u_r^{j+1}(\xi)$, а соответствующее число ℓ через $\ell_r^{j+1}(\xi)$. Максимальное значение индекса j, для которого определены числа $u_r^j(\xi)$ и $\ell_r^j(\xi)$,

будем обозначать через $C_r(\xi)$, а результат применения j-той подстановки к слову ξ, через

/3.22/ $\eta_r^j(\xi) = F^{u_r^j(\xi)}_{l_r^j(\xi)}(\xi)$

Таким образом в i-том шагу каждого последовательного вывода имеются $C_{r_i}(\xi_i)$ различные возможности для продолжения данного вывода. Это положение иллюстрируется деревом последовательных выводов /рис. 3/, узлы которого соответствуют словам ξ_i, выводимым последовательно из слова ξ, а ветви – различным возможностям последовательного продолжения вывода от данного слова ξ_i. Ветви характеризуются вышеопределенными числами $u_{r_i}^j(\xi_i), l_{r_i}^j(\xi)$. Число таких ветвей будет $C_{r_i}(\xi_i)$.

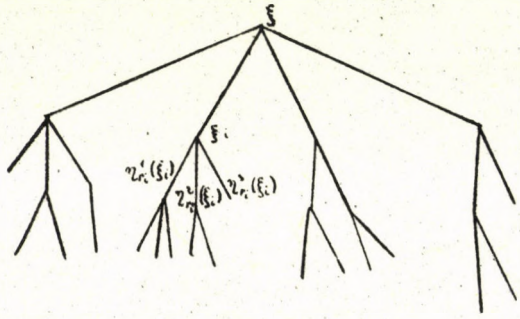


Рис. 3

4. В силу теоремы 4 каждое слово, выводимое из слова ξ, соответствует некоторому узлу дерева последовательных выводов. С другой стороны, очевидно, что все узлы этого дерева могут соответствовать только элементам множества $M_r(\xi)$. Следовательно, теоремы 2 и 3, ограничивающие количество элементов этого множества, ограниничивают и размеры дерева. Таким образом решение задачи анализа – т.е. ответ на вопрос, что выводимо ли слово η из слова ξ – сводится к систематическому просмотру этого дерева, пока не найдется в нем узел, соответствующий искомому слову η.

Этот просмотр осуществляется следующим образом /см. рис. 4/:

1/ Просмотр начинается словом $\xi = \xi_0$, при $r = r_0 = 0$.

2/ В очередном слове ξ_i ищется первое вхождение некоторого слова α_k, заканчивающегося правее от символа порядка r слова ξ_i

3/ Если найдется такое вхождение, заканчивающееся на символе порядка $s > r$ слова ξ_i, то выполняется эта подстановка $(\alpha_k \to \beta_k)$ и будет

/3.23/ $\xi_{i+1} = F_k^s(\xi_i)$, $S_{i+1} = S$, $K_{i+1} = K$.

4/ Если $\xi_{i+1} = \eta$, то поиск закончен, а если нет, то продолжается шагом 2/ со значениями $i = i+1$, $r = s - |\alpha_k|$.

5/ Если при выполнении шага 2/ искомое вхождение не найдется, или каким-то другим образом можно показать, что

/3.24/ $\eta \notin U_r(\xi_i)$,

то вывод не продолжается, а вернемся к слову ξ_{i-1}, и в шаге 6/ рассматривается следующая применимая подстановка к этому слову.

Если $i = 0$, то поиск закончен безрезультатно.

6/ Сперва ищется в слове $\S_i$ вхождение слова a_κ, при $\kappa > \kappa_i$, заканчивающегося на символе порядка S_i этого слова. Если найдется такое вхождение, то вернемся со значением $i = i-1$ к шагу 3/, а если нет, то для поиска дальнейших вхождений, заканчивающихся на некотором символе порядка S при $S > S_i$, - к шагу 2/.

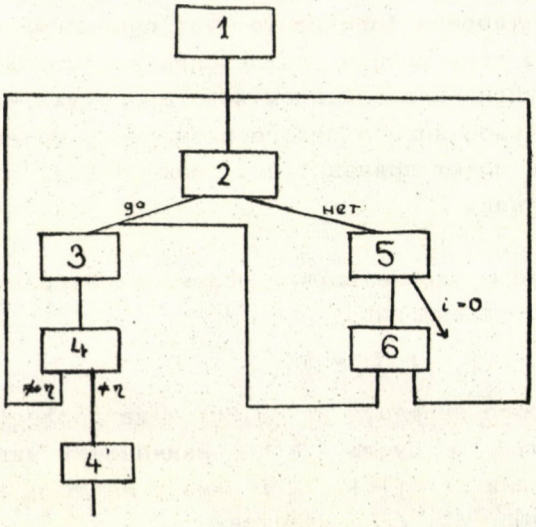


Рис. 4

5. Легко видеть, что последовательность выводов и выбранный метод упорядочения возможных применений подстановок обеспечит то, что ни одна из возможностей не пропускается этим алгоритмом.

Действительно, в шаге 3/ всегда применяется подстановка

/3.25/ $\S_{i+1} = F^{u_{r_i}^{j}(\S)}_{\ell_{r_i}^{j}(\S)}(\S)$

с очередным значением индекса j: в случае первого обращения к этому слову /после шага 4// ищется первое вхождение, заканчивающееся правее от символа порядка r_i, а после j-того - безуспешного - применения некоторой подстановки к этому слову, после 5-ого шага, в шагах 6/ и 2/ действительно ищется /3.25/ для $j = j+1$.

6. Если задача заключается в нахождении не только одного, а <u>всех</u> выводов слова η из слова $\S$, то вместо завершения поисков при выходе из шага 4/, при $\S_{i+1} = \eta$ выполняется дополнительный шаг 4'/, запоминающий найденный вывод, и после этого поиск продолжается как обычно, для нахождения других выводов того же слова η.

Так как алгоритм выполняет полный просмотр дерева последовательных выводов, таким образом все такие выводы найдутся.

7. Тривиальным признаком применимости шага 5/ нашего алгоритма является невозможность применения ни одной подстановки к очередному слову $\S$. В этом случае действительно множество $M_r(\S_i)$ оказывается пустым и так справедливость /3.24/ тривиальна.

Иногда можно предсказать и до появления такого тривиального "тупика" то, что данный вывод не может содержать искомого слова η. Следующая теорема может оказаться полезной для этой цели.

Теорема 5

Если для слова ξ и числа r найдется число v, такое что

/3.26/ $r < v < u_r^*(\xi)$ $\qquad (0 \le r \le |\xi|)$.

и для любой пары чисел k и d

/3.27/ $[\xi]_{v-d}^{v} \neq [\alpha_k]^d$ $\qquad (1 \le d \le |\alpha_k|)$.

то любое слово $\eta \in U_r^*(\xi)$ удовлетворяет равенству

/3.28/ $[\eta]^v = [\xi]^v$.

Другими словами эта теорема формулирует тот очевидный факт, что если в течение поиска следующей возможности применения некоторой подстановки найдется символ, не являющийся продолжением некоторого уже начатого вхождения одного из слов α_k или началом нового такого вхождения, то на этот символ уже ни одна из подстановок не может применяться. Ниже дается и формальное доказательство этого утверждения.

Доказательство:

Пусть слово ξ и число r удовлетворяют условиям /3.26/ и /3.27/. Рассматривается r -последовательный вывод

/3.29/ $F: \xi = \xi_0 \vdash \xi_1 \vdash \ldots \vdash \xi_{t-1} \vdash \xi_t = \eta$

Для первой подстановки этого вывода, из /3.26/ и из последовательности следует, что $v < U_r^*(\xi) \le s_1$. Пусть $i \ge 1$ наименьший индекс для которого $r_i = s_i - |\alpha_{k_i}| < v$, но для $1 \le j \le i-1$ имеет место $r_j \ge v$. Для этих значений индекса j с помощью /1.7/ получается

/3.30/ $[\xi]^v = [\xi_1]^v = \ldots = [\xi_i]^v$ $\qquad (1 \le j \le i-1)$

Предположим, что для слова η утверждение теоремы /3.28/ не имеет места. Из /3.30/ следует, что для этого необходимо существование индекса i с вышеуказанным свойством. Далее, покажем, что в этом случае для $i \ge 2$ тоже имеет место $v < s_i$. Действительно, из $v \ge s_i$, согласно последовательности вывода следовало бы $v \ge s_i > r_{i-1}$, что противоречит определению числа i.

Таким образом получается /см. рис. 5/

/3.31/ $r_i < v < s_i$

Из этого уже следует нарушение условия /3.27/, так как

/3.32/ $\xi_{i-1} = [\xi_{i-1}]^{r_i} \alpha_{k_i} [\xi_{i-1}]_{s_i} = [\xi]^{r_i} [\xi]_{r_i}^{v} [\alpha_{k_i}]_{v-r_i} [\xi_{i-1}]_{s_i}$

откуда получается

/3.33/ $[\xi]_{r_i}^{v} = [\alpha_{k_i}]^{v-r_i}$

что для $d = v - r_i$ и $k = k_i$

действительно противоречит условию /3.27/, и так теорема доказана.

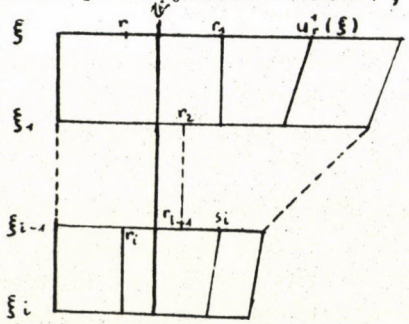


Рис. 5

Согласно только что доказанной теореме, если в некотором выводе получается слово $\xi \cdot \xi_i$, такое, что при выполнении шага 2/ нашего алгоритма, до нахождения следующей применимой подстановки найдется число ν, удовлетворяющее условиям теоремы 5, то можно утверждать, что в дальнейшем ходе вывода первые ν символов слова ξ_i уже остаются неизменными. Таким образом, если искомое слово η отличается, в первых ν символах от слова ξ_i, то /3.24/ можно считать доказанным и так перейти на применение шага 5/.

IV. РЕАЛИЗАЦИЯ АЛГОРИТМА АНАЛИЗА НА ЭВМ

1. В практической реализации на ЭВМ алгоритма анализа, изложенного в главе III, самая большая часть времени может быть затрачена на выполнение шага 2/, осуществляющего поиск возможных применений подстановок, т.е. поиск вхождения некоторого из слов α_κ в соответствующей части слова ξ_i. Этот же шаг должен проверить выполнение условий теоремы 5 для того, чтобы в случае надобности во время перейти к выполнению шага 5/ алгоритма. Обе задачи связаны с поиском вхождения некоторых частей слов α_κ в слово ξ_i.

Для этого ниже излагается алгоритм для распознавания некоторых простых "свойств" слов. Эти свойства связаны с наличием или отсутствием вхождений некоторых из данного набора слов α_κ /или их сегментов/ в данное слово. Излагаемый алгоритм предполагается быть удобным для машинной реализации, так как он основан на операции над булевыми векторами, реализуемыми в ЭВМ машинными словами.

Булевым вектором будем называть набор P логических величин "истина" и "ложь", обозначаемых через 1 и 0 соответственно. По аналогии с машинными словами, координаты этих векторов называем разрядами. Булевые векторы обозначаются через прописные латинские буквы и i-тый разряд вектора X обозначается через X^i. Логические операции конъюнкции /$\wedge$/ и дизъюнкции /$\vee$/ определяются поразрядно, в смысле обычных машинных операций логического умножения и сложения. Арифметические операции определяются как операции над двоичными числами, соответствующими данному вектору. Так, например умножение на 2^{-1} означает сдвиг булевого вектора на один разряд направо. Вектор, содержащий только 0-и, обозначается через $\underline{0}$.

Вводим специальное обозначение: через $X(k)$ обозначается вектор, содержащий единицу только в том разряде, где в векторе X стоит k-тая /считая слева/ единица. Так, если вектор X содержит всего d единиц, то $X = \bigvee_{k=1}^{d} X(k)$

2. Алгоритм непосредственно вытекает из следующей теоремы /в другой формулировке см. Домелки [13] и [14]/.

Теорема 6

Для любой системы формальных подстановок F можно указать число p и булевые векторы длины p

/4.1/ $B_1, B_2, \ldots, B_n, U, V$

такие, что для любого слова

/4.2/ $\xi = \omega_{i_1} \omega_{i_2} \ldots \omega_{i_{|\xi|}}$

векторы Q определенные рекурсией,

/4.3/ $$\begin{cases} Q_0 = \underline{0} \\ Q_t = (Q_{t-1} \cdot 2^{-1} \vee U) \wedge B_{i_t} & (1 \le t \le |\xi|) \end{cases}$$

обладают следующими свойствами

а/ $Q_t \wedge V^{(k)} \neq \underline{0}$ тогда и только тогда, если

/4.4/ $[\xi]_{t-|\alpha_\kappa|}^{t} = \alpha_\kappa$ $(1 \le k \le p)$

б/ $Q_t = \underline{0}$ тогда и только тогда, если для любых чисел k и j

/4.5/ $[\xi]_{t-j}^{t} \neq [\alpha_\kappa]^j$ $(1 \le k \le p,\ 1 \le j \le |\alpha_\kappa|)$

Доказательство:

Пусть сперва рассматривается система, содержащая только одну подстанов-

ку, т.е. $t=1$, $\alpha_1=\alpha$ и $\beta_1=\beta$. В этом случае $p=|\alpha|$, и векторы /4.1/, которые будем называть <u>изображением</u> системы, определяются следующим образом:

$$B_i^j = \begin{cases} 1, & \text{если } [\alpha]_{j-1}^{j} = \omega_i \\ 0, & \text{во всех остальных случаях} \end{cases}$$

/4.6/ $U = /1,0 \ldots 0,0/$

$V = /0,0 \ldots 0,1/$

Покажем, что определенные таким образом векторы выполняют утверждения теоремы.

Пусть для каждого слова ξ, написанного в форме /4.2/ определяется булева функции $g(t,j)$ для $1 \le t \le |\xi|$ и $1 \le j \le p$ следующим образом:

/4.7/ $$g(t,j) = \begin{cases} 1, & \text{если } [\xi]_{t-j}^{t} = [\alpha]^{j} \\ 0, & \text{в других случаях} \end{cases}$$

Легко видеть, что

/4.8/ $$g(t,j) = B_{i_{t-j+1}}^{1} \wedge B_{i_{t-j+2}}^{2} \wedge \ldots \wedge B_{i_t}^{j} = \bigwedge_{r=1}^{j} B_{i_{t-j+r}}^{r}.$$

/здесь предполагается $B_{i_0} = B_{i_{-1}} = B_{i_{-2}} = \ldots = Q$ /

Отсюда для $j \ge 1$ получается

/4.9/ $$g(t,j) = \left(\bigwedge_{r=1}^{j-1} B_{i_{t-j+r}}^{r}\right) \wedge B_{i_t}^{j} = g(t-1,j-1) \wedge B_{i_t}^{j}$$

Определяем для $1 \le t \le |\xi|$ булевые векторы $Q(t)$ следующим образом:

/4.10/ $$Q(t) = (g(t,1), g(t,2), \ldots, g(t,p))$$

и покажем, что они совпадают с векторами Q_t, определенными рекурсией /4.3/.

Действительно, $g(0,j)=0$ для любого числа j и так

/4.11/ $$Q(0) = (g(0,1), g(0,2), \ldots, g(0,p)) = Q = Q_0$$

Если далее положить $Q(t-1) = Q_{t-1}$, то

$$Q_t = (Q_{t-1} \cdot 2^{-1} \vee U) \wedge B_{i_t} =$$

/4.12/ $$= (1, g(t-1,1), g(t-1,2), \ldots, g(t-1,p-1)) \wedge (B_{i_t}^{1}, B_{i_t}^{2}, B_{i_t}^{3}, \ldots\ldots, B_{i_t}^{p})$$

и так согласно /4.9/

/4.13/ $$Q_t = (g(t,1), g(t,2), g(t,3), \ldots, g(t,p)) = Q(t)$$

что и требовалось доказать.

Отсюда из /4.7/ получается

а/ $Q_t \wedge V^{(k)} = Q(t) \wedge V \neq Q$ тогда и только тогда, если $g(t,p)=1$ т.е. если имеет место /4.4/

б/ $Q_t = Q(t) = Q$ тогда и только тогда, если $g(t,j)=0$ для всех , то есть если /4.5/ справедливо.

Таким образом теорема доказана на случай одной подстановки.

В случае $\ell \ge 2$ пусть будет $p = \sum_{\mu=1}^{\ell} |\alpha_\mu|$ и изображение системы определяется как записанные подряд векторы, определенные согласно /4.6/ для отдельных подстановок $(\alpha_\kappa \to \beta_\kappa)$. Покажем, что в этом случае для отдельных векторов $Q_\kappa(t)$ определенных согласно /4.10/ и /4.7/ для отдельных подстановок, имеет место

/4.14/ $$Q_t = Q_1(t)\, Q_2(t) \ldots\ldots Q_\ell(t).$$

Для этого необходимо показать, что при вычислении вектора Q_t согласно формуле /4.3/ все "отрезки" вектора, соответствующие отдельным подстановкам, вы-

числяются независимо друг от друга. В самом деле, операции $\wedge$ и $\vee$, принимающие участие в формуле /4.3/ выполняются поразрядно. Таким образом, только через операции сдвига направо вычисление одного отрезка может влиять на другое, т.е. первый разряд каждого отрезка в векторе $Q_{t-1}\cdot 2^{-1}$ совпадает с последним разрядом предыдущего отрезка в векторе Q_{t-1}. Но после операции сдвига сразу следует операция дизъюнкции с вектором U, содержащим единицы именно в этих первых разрядах отрезков. Таким образом, значение первых разрядов каждого отрезка в векторе $Q_{t-1}\cdot 2^{-1}\vee U$ уже оказывается независимым от предыдущих отрезков, и так формула /4.14/ доказана. Отсюда доказательство утверждений а/ и б/ получается так же, как и в случае $t=1$

3. В реализации алгоритма анализа теорема 6 применяется следующим образом:

1/ Процесс начинается словом ξ при $r=0$ и $Q_0=\Omega$.

2/ Исходя из значения Q_r, для каждого очередного символа ω_{i_t} при $t\geq r+1$ вычисляется значение вектора Q_t по формуле /4.3/ Эти значения вместе с символами ω_{i_t} запоминаются в некотором массиве S.

3/ Если $Q_t\wedge V\neq\Omega$, то ищется наименьшее число k, для которого $Q_t\wedge V^{(k)}\neq\Omega$ и выполняется подстановка $(\alpha_k\to\beta_k)$. Информация о выполненной подстановке тоже запоминается в массиве S. Эта информация содержит

а/ номер подстановки k;

б/ индекс первого символа левой стороны подстановки, т.е. символа $\omega_{i_{t-|\alpha_k|+1}}$ в массиве S;

в/ индекс информации о последней выполненной подстановке в массиве S.

4/ Если в результате подстановки не получилось искомое слово η, то из массива S восстанавливается значение вектора $Q_{t-|\alpha_k|}$, и шаг 2/ продолжается со значением $r=t-|\alpha_k|$, т.е. вычислением векторов Q для только что подставленных символов слова β_k.

Если искомое слово η найдено, то поиск закончится /или выполняется шаг 4', согласно п. 6, гл. III/.

5/ Если $Q_t=\Omega$, то последняя выполненная подстановка устраняется /если такого нет, то происходит безрезультатное окончание процесса/ и из массива S с помощью информации а/, б/ и в/ восстанавливается бывшее значение индекса t во время применения этой подстановки, и значение соответствующего вектора Q_t.

6/ Если $Q_t\wedge(V-V^{(k)})\neq\Omega$, то это означает, что на t-ом символе слова ξ заканчивается еще одна подстановка. В этом случае, для выполнения этой подстановки перейдем к шагу 3/.

Если с другой стороны такой подстановки нет, то продолжается поиск применяемой подстановки переходом к шагу 2/ со значением $r=t$.

Легко видеть, что самые трудоемкие части алгоритма действительно упрощаются применением теоремы 6, позволяющей выполнить проверку целого ряда логических условий с помощью операций над булевыми векторами.

4. В доказательстве теоремы 6 был использован простейший метод построения изображения системы. Иногда это может быть сделано более экономично с точки зрения требуемой памяти.

Легко видеть, например, что если подстановки $(\alpha_k\to\beta_k)$ и $(\alpha_l\to\beta_k)$ различаются только в одном символе левой части, т.е. $\beta_k=\beta_l$ и

/4.15/ $\alpha_k=\varphi\omega_r\tau$, $\alpha_l=\varphi\omega_s\tau$,

то в векторах /4.1/ можно обойтись без отрезка, соответствующего подстановке $(\alpha_l\to\beta_l)$, если не только вектор β_r, но и вектор β_s тоже содержит единицу в

соответствующем разряде отрезка, соответствующего подстановке $(\alpha_k \to \beta_k)$.

Действительно, в этом случае, при вхождении слов α_k и α_ℓ одинаково получается единица в соответствующем разряде слова $Q_t \wedge V^{(h)}$, и, так как $\beta_\ell = \beta_k$, выполняется нужная подстановка.

В общей форме этот факт может быть сформулирован следующим образом:

Пусть правые части некоторых подстановок совпадают, и левые части этих подстановок имеют общую длину p и их можно устроить в форме таблицы

/4.16/
$$T = \left\| \begin{array}{cccc} \omega_{11} & \omega_{21} & \dots & \omega_{p1} \\ \omega_{12} & \omega_{22} & \dots & \omega_{p2} \\ \vdots & \vdots & & \vdots \\ \omega_{1h_1} & \omega_{1h_2} & \dots & \omega_{ph_p} \end{array} \right\|$$

таким образом, что любое сочетание элементов из различных столбцов

/4.17/ $\omega_{1i_1} \; \omega_{2i_2} \dots \omega_{pi_p} \qquad (1 \le i_j \le h_j)$

является левой частью одной подстановки из данной группы. В этом случае будем говорить, что группа подстановок <u>соответствует таблице</u> T. Из вышесказанного следует, что если в изображении системы этой группе соответствует один отрезок, определенный соотношением

/4.18/
$$B_i^1 = \begin{cases} 1, \text{ если для некоторого числа } \ell \text{ имеет} \\ \quad \text{место } \omega_i = \omega_{i\ell} \\ 0, \text{ в остальных случаях} \end{cases}$$

то вхождение левой части любой из подстановок данной группы в некотором слове ξ обнаруживается методом, изложенным в теореме 6.

Следует отметить, что при вышеуказанном методе составления изображения системы, количество подстановок в группе, соответствующей одной таблице, никак не влияет ни на длину булевых векторов, ни на количество выполняемых операций. Пользуясь этим фактом можно, например, исключить все подстановки вида $(\omega_i \to \omega_j)$ из системы F. Для этого достаточно прибавить к каждой подстановке $(\alpha_k \to \beta_k)$ системы $F \setminus (\omega_i \to \omega_j)$ все подстановки $(\alpha_k' \to \beta_k')$ полученные заменой всех вхождений символа ω_j в слове α_k на символ ω_i. Легко видеть, что если

/4.19/ $\alpha_k = \gamma_1 \omega_j \gamma_2 \omega_j \gamma_3 \dots \gamma_{b-1} \omega_j \gamma_b$,

то группа подстановок $(\alpha_k' \to \beta_k)$ вместе с подстановкой $(\alpha_k \to \beta_k)$ соответствует таблице

/4.20/ $T_k = \| \gamma_1 \overset{\omega_i}{\omega_j} \; \gamma_2 \overset{\omega_i}{\omega_j} \; \gamma_3 \dots \gamma_{b-1} \overset{\omega_i}{\omega_j} \; \gamma_b \|$

и так, это преобразование действительно не увеличит длину изображения системы.

5. В случае больших систем, запоминание булевых векторов, составляющих изображение системы может потребовать очень большого объема памяти ЭВМ. С другой стороны, если эти векторы запоминаются в последовательных машинных словах, то большое количество этих слов может содержать только нули. Поэтому, может оказаться целесообразным запоминать только ненулевые из этих ма-

шинных слов, вместе с серийным номером данного слова внутри булевого вектора. Это, конечно, в некоторой степени усложняет операции над булевыми векторами, написанными в такой форме, но с другой стороны лишние операции над нулями будут сэкономлены. /Подробное описание такого алгоритма, вместе с программой, написанной на языке АЛГОЛ-60, см. в работе Домелки [15]. В программе, написанной для машины УРАЛ-2 /см. п. 5. Введения/ Д. Варга разрабатывал новый метод для запоминания очень длинных булевых векторов, используя для этого "оглавления", состоящиеся из булевых векторов, содержащих единицу в разрядах, соответствующих ненулевым словам длинного булевого вектора.

6. Время выполнения алгоритма анализа в отдельных случаях намного зависит от количества "тупиков" в течение анализа, т.е. от того, сколько раз обнаруживается "бесперспективность" данного вывода, и применяются шаги 5/ и 6/ для устранения последней выполненной подстановки. Если, например, при этом понадобится выбирать другую возможность вместо некоторой подстановки в <u>начале</u> вывода, а обнаруживается такая необходимость - путем получения $Q = Q$ - только в конце вывода, после выполнения целого ряда "правильных" подстановок, то согласно алгоритму эти правильные подстановки тоже должны быть по очереди устранены, пока не найдется действительно устранимая подстановка. Легко видеть, что хотя алгоритм и теперь правильно работает, но в худшем случае количество операций может носить экспоненциальный характер.

Для того, чтобы улучшить положение в таких случаях вводится понятие <u>финальной</u> подстановки. Так назовем подстановку, о которой каким-то образом можно доказать, что любое ее применение является окончательным в том смысле, что применение вместо ее другой подстановки не может привести в дальнейшем к слову, из которого искомое слово выводимо. /В качестве простого примера, легко видеть, что если слово α_k заканчивается на символ ω_i, не входящий ни в другие слова α_l при $l \neq k$, ни в слово η, то подстановку $(\alpha_k \to \beta_k)$ можно считать финальной/.

Таким образом, пока применяются только финальные подстановки, можно выполнить подстановку прямо в последовательности анализируемых символов, без отметки в массиве S, в результате чего, если обнаруживается необходимость устранения последней примененной подстановки, то эти подстановки уже не устраняются. Кроме того, выполнение финальной подстановки должно превратить в финальные все обыкновенные подстановки внутри ее области действия, т.е. эти подстановки тоже должны быть опущены из массива S. /Точное определение финальных подстановок и изложение некоторых их свойств дается в гл. V./

Далее, если цель анализа является какой-то перевод анализируемого текста, то после выполнения финальных подстановок можем обратиться к подпрограммам, осуществляющим перевод некоторых частей текста. Эти подпрограммы будем называть <u>семантическими подпрограммами</u>, и поскольку их характер определяется задачей, в интересах которого производится анализ, в данной работе не будем заниматься формой и содержанием этих подпрограмм.

V. ПРЯМЫЕ ВЫВОДЫ И ФИНАЛЬНЫЕ ПОДСТАНОВКИ

В этой главе даются более точные определения понятий, связанных с упомянутым в п. 6 гл. IV методом ускорения анализа путем уменьшения количества тупиков. Пример в конце главы /п.7/ служит для объяснения введенных понятий.

1. Последовательный вывод

/5.1/ $F: \xi = \xi_0 \vdash \xi_1 \vdash \ldots \vdash \xi_t = \eta,$

где

/5.2/ $\xi_i = F^{s_i}_{K_i}(\xi_{i-1}) \qquad (1 \le i \le t),$

называется прямым относительно подстановки $(\alpha_K \to \beta_K)$, если при поиске вывода согласно алгоритму, изложенному в гл. III, возможность применения этой подстановки ни один раз не пропускается, т.е. если для любого числа i из

/5.3/ $K_i = l^h_{r_{i-1}}(\xi_{i-1})$

следует, что $k \ne l^j_{r_{i-1}}(\xi_{i-1})$ при $1 \le j \le h$.

Вывод называется прямым, если он является прямым относительно всех подстановок данной системы.

Легко видеть, что вывод тогда и только тогда является прямым, если

/5.4/ $K_i = l^1_{r_{i-1}}(\xi_{i-1}) \qquad (1 \le i \le t).$

Таким образом из каждого слова существует только один прямой вывод, каждый шаг которого однозначно определен. Вводится обозначение

/5.5/ $F: \xi \Longrightarrow \eta$

Это означает, что слово η выводимо из слова ξ с помощью прямого вывода.

Если прямой вывод из слова ξ заканчивается, т.е. найдется слово $\xi_t = \eta$, для которого ни одна из подстановок не применима $(C_{r_t}(\xi_t) = 0)$, то это слово обозначается через $\eta = \pi_F(\xi)$. Если прямой вывод не заканчивается, то функция $\pi_F(\xi)$ считается неопределенной. Таким образом, каждая система F однозначно определяет алгоритм π_F, применение которого к любому слову ξ заключается в выполнении шагов прямого вывода, пока это оказывается возможным.

2. Покажем, что это новое понятие алгоритма, определенное через прямые выводы, равносильно известным другим понятиям алгоритмов /как, например, машины Туринга, нормальные алгорифмы, частично-рекурсивные функции/.

Теорема 7

а/ Для любой системы формальных подстановок F в алфавите A существует нормальный алгорифм $\mathcal{A}$ над алфавитом A, что для любого слова $\xi \in S(A)$

/5.6/ $\mathcal{A}(\xi) \simeq \pi_F(\xi)$

б/ Для любого нормального алгорифма $\mathcal{A}$ в алфавите A существует система формальных подстановок F в алфавите $\tilde{A} \supset A$ и символы $\gamma, \delta \in \tilde{A} \setminus A$, такие что для любого слова $\xi \in S(A)$

/5.7/ $\pi_F(\gamma \xi \delta) \simeq \delta\, \mathcal{A}(\xi)\, \gamma$

/Об определении понятий связанных с нормальными алгорифмами см. Марков [8]. В частности, введенное там понятие условного равенства $\simeq$ означает, что если один из объектов, связанных с этим знаком, существует, то и другой должен существовать и они должны быть равными/.

Доказательство:

а/ Пусть δ обозначает букву, не принадлежащую алфавиту A. Построим нор-

мальный алгориям $\mathcal{A}$ в алфавите $A \cup \sigma$, с сокращенно записанной схемой

/5.8.1/ $\alpha_k \sigma \to \sigma \beta_k$ $(1 \le k \le f)$

/5.8.2/ $\sigma \omega_i \to \omega_i \sigma$ $(1 \le i \le n)$

/5.8.3/ $\sigma \to .$

/5.8.4/ $. \to \sigma$

Легко видеть, что применение этого алгорифма на любое слово $\xi \in S(A)$ соответствует прямому выводу /5.1/ в следующем смысле: после первоначального применения формулы /5.8.4/, каждому шагу $\xi - F_{ki}^{s_i}(\xi_{i-1})$ вывода /5.1/ соответствуют следующие шаги применения алгорифма $\mathcal{A}$: выполнить $s_i - r_{i-1}$ раз формулы /5.8.2/, потом применить формулы /5.8.1/ /здесь предполагается $r_0 = 0$/. В результате такой i-той группы шагов получается слово

/5.9/ $\eta_{j(i)} = [\xi_{i-1}]^{r_i} \sigma \beta_{ki} [\xi_{i-1}]_{s_i}$

На это слово опять применяются $s_{i+1} - r_i$ раз формулы /5.8.2/, т.е. символ σ передвигается направо на такое же количество мест и применяется формула $(\alpha_{k_{i+1}} \sigma \to \sigma \beta_{k\,i-1})$ и т.д., пока символ σ не дойдет до конца слова, когда формула /5.8.3/ заканчивает применение алгорифма.

Тот факт, что вывод /5.1/ является прямым, т.е. всегда выполняется первая возможность применения подстановки, соответствует правильному порядку применения формул алгорифма $\mathcal{A}$, и так слова $\eta_{j(i)}$ для $i = 1,2 \ldots t$ соответствуют словам ξ_i вывода /5.1/.

Таким образом результаты применения алгорифма $\mathcal{A}$ и выполнения прямого вывода /5.1/ действительно совпадают.

Замечание: Количество выполнений формул /5.8.2/ в каждой группе уменьшается на 1, если вместо формул /5.8.1/ взять формулы

/5.8.1'/ $\alpha_k \sigma \to [\beta_k]' \sigma [\beta_k]_1$

б/ Пусть $\mathcal{A}$ нормальный алгорифм в алфавите A с сокращенно записанной схемой

/5.10/ $K_i \to \mathcal{A}_i$ $(1 \le i \le m)$

Для $1 \le i \le m$ определяются непересекающиеся "копии"

/5.11/ $A^{(i)} = \{ \omega_1^{(i)}, \omega_2^{(i)}, \ldots, \omega_n^{(i)} \}$

алфавита A, и определяется

/5.12/ $\tilde{A} = A \cup \bigcup_{i=1}^{m} A^{(i)} \cup \{ \gamma, \gamma', \sigma, \sigma', \sigma'', \varepsilon, \varepsilon' \}$

где символы $\gamma, \gamma', \sigma, \sigma', \sigma'', \varepsilon, \varepsilon'$ не принадлежат алфавитам A или $A^{(i)}$

Система F составляется следующим образом:

1/ задаются подстановки, переписывающие слово ξ в алфавит $A^{(1)}$:

$$\gamma \omega_j \to \gamma \omega_j^{(1)} \gamma' \qquad (1 \le j \le n)$$
$$\gamma' \omega_j \to \omega_j^{(1)} \gamma' \qquad (1 \le j \le n)$$
$$\gamma' \sigma \to \sigma$$

2/ для каждой незаключительной формулы $K_\rho \to \mathcal{A}_\rho$ схемы алгорифма $\mathcal{A}$ задается группа подстановок, которые выполняют эту формулу над словом в алфавите $A^{(\rho)}$ и потом переписывают полученное слово в алфавит $A^{(1)}$:

$$K_\rho^{(\rho)} \to \sigma' \mathcal{A}_\rho^{(1)} \gamma$$
$$\omega_j^{(\rho)} \sigma' \to \sigma' \omega_j^{(1)} \qquad (1 \le j \le n)$$
$$\gamma \sigma' \to \gamma$$
$$\gamma' \omega_j^{(\rho)} \to \omega_{(j)}^{(1)} \gamma' \qquad (1 \le j \le n)$$
$$\gamma' \sigma \to \sigma$$

3/ для каждой заключительной формулы $K_q \to \lambda_q$ задается группа подстановок, которые выполняет эту формулу над словом в алфавите $A^{(q)}$ и потом переписывают это слово в первоначальный алфавит A, причем символы γ и δ переменяются местами, что является признаком конца работы:

$$K_q^{(q)} \to \varepsilon' \lambda_q \varepsilon$$
$$\omega_j^{(q)} \varepsilon' \to \varepsilon' \omega_j \qquad (1 \le j \le n)$$
$$\gamma \varepsilon' \to \delta$$
$$\varepsilon \omega_i^{(q)} \to \omega_i \varepsilon \qquad (1 \le j \le n)$$
$$\varepsilon \delta \to \gamma$$

4/ для $i = 1, 2, \ldots, m-1$ задаются подстановки, которые в случае неприменимости i-той формулы на слово в алфавите $A^{(i)}$ переписывают это слово в алфавит $A^{(i+1)}$:

$$\omega_j^{(i)} \delta \to \delta'' \omega_j^{(i+1)} \delta \qquad (1 \le j \le n)$$
$$\omega_j^{(i)} \delta'' \to \delta'' \omega_j^{(i+1)} \qquad (1 \le j \le n)$$
$$\gamma \delta'' \to \gamma$$

5/ Для $i = m$ задаются подстановки, которые соответствуют естественному останову алгорифма $\mathcal{A}$, если ни одна из подстановок не применима. В этом случае выполняются те же самые действия как и в случае 3/:

$$\omega_j^{(m)} \delta \to \varepsilon' \omega_j \gamma \qquad (1 \le j \le n)$$
$$\omega_j^{(m)} \varepsilon' \to \varepsilon' \omega_j \qquad (1 \le j \le n)$$
$$\gamma \varepsilon' \to \delta$$

Нетрудно убедиться в том, что заданные подстановки действительно выполняют действия, приписанные им, и что, согласно правилам прямого вывода, они выполняются в правильном порядке. Строгое доказательство этого факта опускается.

Таким образом, прямой вывод в системе F работает точно так же как и нормальный алгорифм $\mathcal{A}$: возможности выполнения различных формул проверяются по упорядочению формул, и всегда выполняется первое вхождение левой части данной формулы. После выполнения некоторой формулы все начинается сначала, с первой формулы.

Если и последняя формула оказывается неприменимой, или, если выполнена заключительная формула, то вывод заканчивается и получается слово $\delta \mathcal{A}(\xi) \gamma$.

Замечание: Так как формулы алгорифма $\mathcal{A}$ различаются по различным алфавитам, очевидно, что они не должны содержать формулу, левая часть которой пустая. В таком случае следует вместо формулы

/5.13/ $\qquad \longrightarrow \lambda_i$

в схеме алгорифма $\mathcal{A}$, рассматривать формулы

/5.14/ $\qquad \omega_j^{(i)} \longrightarrow \omega_j^{(i)} \lambda_i^{(i)} \qquad (1 \le j \le n)$

3. Универсальность класса алгорифмов, определенных с помощью прямых выводов в системах формальных подстановок, доказанная в теореме 7, дает возможность на решения задачи анализа в ряде случаев с помощью прямых выводов, без просмотра дерева последовательных выводов, как это было делано в алгоритме гл. III.

Действительно, пусть задача анализа системы F относительно множества $T \subseteq S(A)$ является разрешимой в том смысле, что существует нормальный алгоритм $\mathcal{A}$, перерабатывающий слово ξ в непустое слово тогда и только тогда, если $F \cdot \xi \vdash \eta \in T$. /Легко видеть, например, что согласно сказанному в п. 3 гл. II, во всех ограниченных системах с рекурсивной функцией $b_F(\xi)$ задача анализа является разрешимой относительно рекурсивно определенных - в том числе

конечных - множеств T/. В этом случае система F', построенная согласно части б/ теоремы 7 к нормальному алгорифму $\mathcal{A}$, обладает следующим свойством: для любого слова $\xi \in S(A)$ тогда и только тогда будет $F: \xi \vdash \eta \in T$, если

/5.15/ $F': \gamma\xi\delta \Rightarrow \delta\eta'\gamma$, где $\eta' \neq \Lambda$

Если, далее, нормальный алгорифм $\mathcal{A}$ построен так, чтобы непустое слово, в которое перерабатывается слово ξ, содержало некоторое описание одного из выводов из слова ξ, ведущего в некоторый элемент множества T, то построенная таким же образом система F' дает в качестве результата прямого вывода слово η, содержащее описание оригинального вывода в системе F.

Таким образом, задача анализа системы F полностью решается с помощью прямых выводов в системе F'.

Подобное утверждение было доказано в работе Мазуркевича [11] о контекстно-свободных языках и <u>простых</u> выводах, отличающихся от прямых выводов в том, что в них требуется, чтобы на каждом шаге вывода была применима только одна подстановка $(C_{r_i}(\xi_{i-1}) = 1$ для $1 \leq i \leq t)$.

4. Система F', построенная согласно теореме 7 с помощью нормальных алгорифмов, для решения задачи анализа системы F путем прямых выводов, может оказаться слишком громоздкой для практической реализации, и количество выполнимых шагов может быть значительно больше, чем длина вывода в системе F. Применение прямых выводов для решения задачи является более эффективным, если для данной системы F и множества T можно найти такую систему G, которая для любого слова ξ "выбирает" из всех возможных выводов в системе F один, ведущий в множество T /если такой вывод существует/.

Другими словами, от системы G требуется, что если множество $T \cap M_F(\xi)$ непустое, то для одного из его элементов $\eta \in T \cap M_F(\xi)$ имело место $G: \xi \Rightarrow \eta$ причем длина этого прямого вывода соответствовала бы длине вывода слова η в системе F.

Общее решение этой проблемы здесь не дается. В гл. VI описана подобная система для синтаксиса языка АЛГОЛ-60, построенная с использованием некоторых особенностей этого языка, а пример в п. 7 покажет, какими методами можно при этом пользоваться.

В поисках прямого вывода, соответствующего выводу в системе F, те подстановки, относительно которых этот вывод является прямым, не представляют никаких трудностей. Поэтому цель заключается в том, чтобы переделать все подстановки системы в такие.

В связи с этим уточняется введенное в конце гл. IV интуитивное понятие финальной подстановки.

Подстановка $(\alpha_k \rightarrow \beta_k)$ называется <u>финальной</u> относительно множества $T \subseteq S(A)$, если для любого слова $\xi \in S(A)$ либо не существует вывода $F: \xi \vdash \eta \in T$ ведущий в множество T, либо хотя бы один из этих выводов является прямым относительно этой подстановки.

Смысл этого определения объясняется с помощью дерева последовательных выводов /см. п. 3 гл. III/: если в дереве, принадлежащем к слову ξ, найдутся пути, соответствующие выводам, ведущим в множество T, то один из этих выводов должен быть прямым относительно подстановки $(\alpha_k \rightarrow \beta_k)$, т.е. ветви, отходящие <u>слева</u> от этого пути не должны соответствовать этой подстановке /см. рис. 6/. Пути, ведущие в множество T обозначены толстой линией и черточкой отмечены ветви, которые не должны соответствовать подстановке $(\alpha_k \rightarrow \beta_k)$

Легко видеть, что это определение соответствует старому понятию финальности: если ищется первый прямой вывод, ведущий в множество T и найдется

возможность применения подстановки $(\alpha_k \to \beta_k)$, то мы можем ее сразу выполнить, так как из финальности следует, что если вместо ее выполняется другая /стоящая по упорядочению возможностей после нее/ подстановка, то вывод либо перестает быть прямым относительно подстановки $(\alpha_k \to \beta_k)$, либо не должен быть единственным ведущим в множество T.

Иногда, если ищутся <u>все</u> выводы, ведущие в множество T, понадобится другое понятие финальности подстановок.

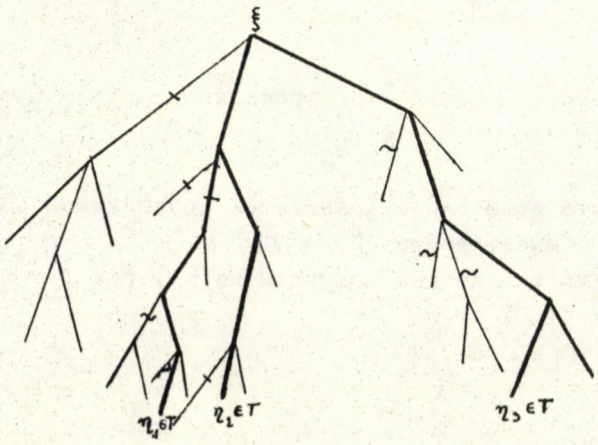


Рис. 6

Подстановка $(\alpha_k \to \beta_k)$ называется <u>финальной в сильном смысле</u> относительно множества $T \subseteq S(A)$, если <u>все</u> выводы, ведущие в множество T являются прямыми относительно этой подстановки. /На рис. 6 это означает, что ветви, отмеченные волнистой линией $(\sim)$, тоже не должны соответствовать подстановке $(\alpha_k \to \beta_k)$/.

Легко видеть, что подстановки, финальные в этом смысле, действительно удовлетворяют более строгим требованиям, чем раньше: из финальности подстановки в сильном смысле следует ее финальность в обыкновенном смысле.

5. Для распознавания финальных подстановок ниже дается необходимое и достаточное условие финальности подстановки в сильном смысле, которое и является достаточным условием ее финальности в обыкновенном смысле. Заметим, что необходимое условие финальности служило бы только для распознавания нефинальности некоторой подстановки, но в этом всегда можно убедиться и с помощью соответствующего примера также.

Условие задается с помощью понятия продолжимости слов:

Слово $\xi \in S(A)$ называется <u>продолжимым</u> порядка r относительно множества $T \subseteq S(A)$, если существуют слова ρ, τ, η такие, что

/5.16/ $\quad F : \rho\,\xi\,\tau \models \eta \in T,$

и первым шагом этого последовательного вывода /если его длина > 0/ является применение подстановки $\xi_1 = F_k^s(\rho\,\xi\,\tau)$ где $s > |\rho| + r$.

Другими словами это означает, что для слова ξ существует такой контекст, из которого выводим некоторый элемент множества T, причем этот вывод начинается "правее" от символа порядка r слова ξ.

Пусть $C_F(\xi, r, T)$ есть логическое отношение, имеющее следующее рекурсивное определение:

$C_F(\xi, r, T)$ имеет место тогда и только тогда, если выполняется одно из следующих условий /здесь p, q неотрицательные числа, а $(\alpha_\ell \longrightarrow \beta_\ell)$ подстановка системы F /:

1/ $[\xi]_p^q = \alpha_\ell$, при $q > r, p > 0$ и $C_F([\xi]^p \beta_\ell [\xi]_q, p, T)$

2/ $[\xi]^q = [\alpha_\ell]_p$, при $q > r$ и $C_F(\beta_\ell [\xi]_q, 0, T)$

3/ $[\xi]_p = [\alpha_\ell]^q$, при $p > 0, q < |\alpha_\ell|$ и $C_F([\xi]^p \beta_\ell, p, T)$

4/ $\xi = [\alpha_\ell]_p^q$, при $q < |\alpha_\ell|$ и $C_F(\beta_\ell, 0, T)$

5/ $\xi = [\eta]_p^q$ при $\eta \in T$

Теорема 8

Слово ξ тогда и только тогда является продолжимым порядка r относительно множества T, если имеет место $C_F(\xi, r, T)$.

Доказательство:

а/ Сперва покажем, что если слово ξ является продолжимым порядка относительно множества T, то имеет место $C_F(\xi, r, T)$

Доказательство проводится методом индукции по t, где t длина вывода /5.16/.

Если $t = 0$, то $\gamma \xi \tau = \eta \in T$ и так

/5.17/ $\xi = [\eta]_{|\gamma|}^{|\gamma| + |\xi|}$,

т.е. выполняется условие 5/.

Предположим, что для всех выводов длины $t' < t$ наше утверждение уже доказано. Тогда ищем в выводе /5.16/ наименьшее число i, для которого $r_i < |\gamma| + |\xi|$. Если такого числа нет, то все подстановки вывода выполняются "правее" от слова $\gamma\xi$, и так согласно /1.7/ получается $\eta = \gamma \xi \tau$, откуда следует /5.17/ и, следовательно, условие 5/.

Если найдется такое число i, то опять из /1.7/ получается

/5.18/ $\gamma \xi = [\xi_0]^{|\gamma| + |\xi|} = [\xi_1]^{|\gamma| + |\xi|} = \ldots = [\xi_{i-1}]^{|\gamma| + |\xi|}$

и, поскольку от слова ξ_i длина вывода уже меньше, чем t то - согласно индуктивному допущению - для любого сегмента $[\xi_i]_a^b$ слова ξ /при $r_i \geq a$ / имеет место

/5.19/ $C_F([\xi_i]_a^b, r_i - a, T)$,

так как из последовательности вывода следует, что $s_{i+1} > r_i$ и так вывод из слова ξ_i начинается правее от символа порядка r_i слова ξ_i, т.е. от символа порядка $r_i - a$ данного сегмента. Пусть $k_i = \ell$, тогда в расположении выполняемой подстановки $(\alpha_\ell \longrightarrow \beta_\ell)$ имеются следующие возможности:

1/ $|\gamma| + r < s_i \leq |\gamma| + |\xi|, \quad r_i > |\gamma|$

2/ $|\gamma| + r < s_i \leq |\gamma| + |\xi|, \quad r_i \leq |\gamma|$

3/ $|\gamma| + |\xi| < s_i, \quad r_i > |\gamma|$

4/ $|\gamma| + |\xi| < s_i, \quad r_i \leq |\gamma|$

Во всех этих четырех случаях можно показать выполнение соответствующего условия в определении отношения $C_F(\xi, r, T)$ используя при этом формулу /5.19/, имеющую место согласно индуктивному допущению.

В этом можно убедиться с помощью таблицы А, в которой для каждого случая задаются греницы сегмента a и , что обеспечит применимость соответствующего условия:

б/ Покажем, что если имеет место $C_F(\xi, r, T)$, то слово ξ является продолжимым порядка r относительно множества T. Если имеет место $C_F(\xi, r, T)$, то по определению этого отношения существует последовательность слов и чисел

/5.20/ $\eta_0, d_0, \eta_1, d_1, \ldots, \eta_{u-1}, d_{u-1}, \eta_u = \xi, d_u = r$

такая, что $C_F(\eta_0, d_0, T)$ имеет место по условию 5/ /т.е. $\eta_0 = [\eta]^{\gamma}_{\rho}$, при $\eta \in T$, а для $1 \le i \le u-1$ из $C_F(\eta_i, d_i, T)$ следует $C_F(\eta_{i+1}, d_{i+1}, T)$ по одному из условий 1/-4/.

Таким образом, для $u = 0$ продолжимость слова ξ очевидна, а если предположим, что утверждение уже доказано для всех слов ξ и чисел r , для которых длина последовательности /5.20/ для доказательства отношения $C_F(\xi, r, T)$ меньше чем u, то из этого следует продолжимость слова η_{u-1}, т.е. существуют слова ρ, τ η и числа s, k , для которых $\rho\eta_{u-1}\tau \in T$ или

/5.21/ $F: \rho\eta_{u-1}\tau \vdash F^s_k(\rho\eta_{u-1}\tau) \models \eta \in T,$

где $s > |\rho| + d_{u-1}$

Пусть отношение $C_F(\eta_u, d_u, T)$ следует из $C_F(\eta_{u-1}, d_{u-1}, T)$ согласно одному из условий 1/-4/.

С помощью таблицы В для каждого случая задаются слова ρ', τ' и число s' - такие, что

/5.22/ $F: \rho'\eta_u\tau' \vdash F^{s'}_{\ell}(\rho'\eta_u\tau') = \rho\eta_{u-1}\tau,$

причем $s' > |\rho'| + d_u$ и $s' - |\alpha'| < s$

где это последнее неравенство обеспечит последовательную продолжаемости вывода /5.22/ выводом /5.21/.

Таким образом продолжимость слова $\xi = \eta_u$ порядка $r = d_u$ относительно множества T доказана.

С помощью теоремы 8 можно легко убедиться в непродолжимости относительно данного множества слов

/5.23/ $\xi_1, \quad \xi_2, \ldots, \xi_h$

порядка

/5.24/ $r_1, \quad r_2, \ldots, r_h$

соответственно, в том случае, если ни одно из слов /5.23/ не выполняет условие 5/, а применение условий 1/-4/ на любое слово ξ_i и число r_i всегда ведет к некоторому отношению $C_F(\xi_j, r_j, T)$ где ξ_j, r_j тоже принадлежат к набору слов /5.23/ и чисел /5.24/ соответственно.

В этом случае действительно ни для одного отношения $C_F(\xi_i, r_i, T)$ не может существовать последовательность типа /5.20/ и поэтому слова /5.23/ не могут быть продолжимыми порядка соответствующих чисел из /5.24/ относительно множества T.

6. Понятие продолжимости слов и соответствующее отношение $C_F(\xi, r, T)$ дают возможность для распознавания финальности подстановок относительно множества T.

Теорема 9

Для того, чтобы подстановка $(\alpha_k \to \beta_k)$ была финальной в сильном смысле относительно множества T необходимо и достаточно выполнение следующих условий:

А/ $\neg C_F(\alpha_k, |\alpha_k|, T)$

Таблица А

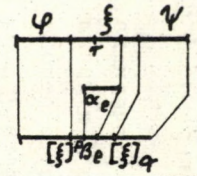
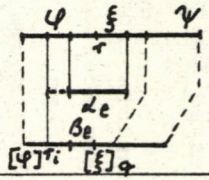
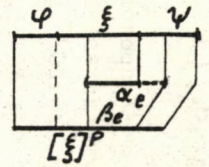
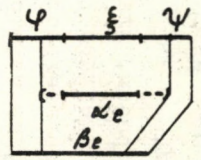
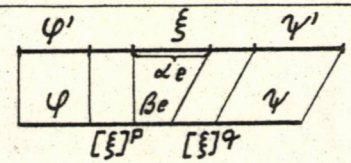
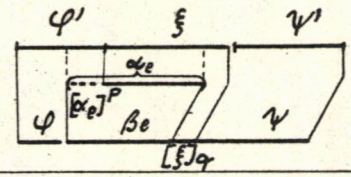
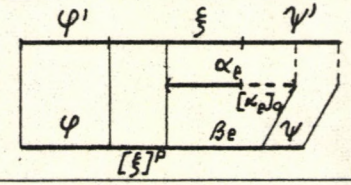
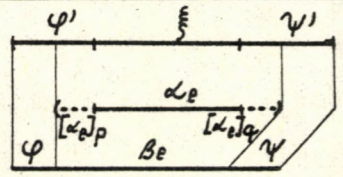
№ случая	α_e	p	q	ξ_i	a	b	
1,)	$[\xi_{i-1}]_{r_i}^{s_i} = [\xi]_{r_i-\lvert\varphi\rvert}^{s_i-\lvert\varphi\rvert}$	$r_i-\lvert\varphi\rvert>0$	$s_i-\lvert\varphi\rvert>\tau$	$\varphi[\xi]^p\beta_e[\xi]_q\psi$	$\lvert\varphi\rvert$	$\lvert\varphi\rvert+p+\lvert\beta_e\rvert+q$	
2,)	$[\varphi]_{r_i}[\xi]^{s_i-\lvert\varphi\rvert}$	$\lvert\varphi\rvert-r_i$	$s_i-\lvert\varphi\rvert>\tau$	$[\varphi]^{r_i}\beta_e[\xi]_q\psi$	r_i	$r_i+\lvert\beta_e\rvert+q$	
3,)	$[\xi]_{r_i-\lvert\varphi\rvert}[\psi]^{s_i-\lvert\varphi\rvert-\lvert\xi\rvert}$	$r_i-\lvert\varphi\rvert>0$	$\lvert\xi\rvert+\lvert\varphi\rvert-r_i<$ $<s_i-r_i=\lvert\alpha_e\rvert$	$\varphi[\xi]^p\beta_e[\psi]_{s_i-\lvert\varphi\rvert-\lvert\xi\rvert}$	$\lvert\varphi\rvert$	$\lvert\varphi\rvert+p+\lvert\beta_e\rvert$	
4,)	$[\varphi]_{r_i}\xi[\psi]^{s_i-\lvert\varphi\rvert-\lvert\xi\rvert}$	$\lvert\varphi\rvert-r_i$	$\lvert\xi\rvert+\lvert\varphi\rvert-r_i<$ $<s_i-r_i=\lvert\alpha_e\rvert$	$[\varphi]^{r_i}\beta_e[\psi]_{s_i-\lvert\varphi\rvert-\lvert\xi\rvert}$	r_i	$r_i+\lvert\beta_e\rvert$	

Таблица Б

№ случая	φ'	ψ'	S'	$S'-\lvert\alpha_e\rvert$	
1.)	φ	ψ	$\lvert\varphi\rvert+q$	$\lvert\varphi\rvert+p$	
2.)	$\varphi[\alpha_e]^p$	ψ	$\lvert\varphi\rvert+p+q$	$\lvert\varphi\rvert$	
3.)	φ	$[\alpha_e]_q\,\psi$	$\lvert\varphi\rvert+p+\lvert\alpha_e\rvert$	$\lvert\varphi\rvert+p$	
4.)	$\varphi[\alpha_e]^p$	$[\alpha_e]_q\,\psi$	$\lvert\varphi\rvert+\lvert\alpha_e\rvert$	$\lvert\varphi\rvert$	

В/ если $\alpha_k = [\alpha_\ell]_p$, при $\ell > k$, $0 \le p \le |\alpha_\ell|$,

то $\neg\, C_F(\beta_\ell, 0, T)$

В/ если $\alpha_\ell = [\alpha_k]_p$, при $\ell > k$, $0 \le p \le |\alpha_k|$,

то $\neg\, C_F([\alpha_k]^p \beta_\ell, p, T)$

Доказательство:

Необходимость этих условий очевидно следует из определения финальности подстановки в сильном смысле: во всех трех случаях из продолжимости соответствующего слова следовало бы существование вывода, ведущего в множество T, который не является прямым относительно подстановки $(\alpha_k \to \beta_k)$. Таким образом, эта подстановка не будет финальной в сильном смысле относительно множества T.

Для доказательства достаточности условия предположим, что подстановка не является финальной в сильном смысле относительно множества T. Тогда существует вывод $F: \xi \vdash \eta \in T$, для i-того шага которого

/5.25/ $\xi_i = F_{ki}^{s_i}(\xi_{i-1})$

где $k_i = \ell^n_{r_{i-1}}(\xi_{i-1})$ и $k = \ell^j_{r_{i-1}}(\xi_{i-1})$

при $1 \le j < n$.

Это означает, что для $u = u^j_{r_{i-1}}(\xi_{i-1})$ и

/5.26/ $\rho = [\xi_{i-1}]^{u-|\alpha_k|}$, $\gamma = [\xi_{i-1}]_u$

имеет место $\xi_{i-1} = \rho\, \alpha_k\, \gamma$.

С другой стороны, $\xi_{i-1} = [\xi_{i-1}]^{r_i}\, \alpha_{ki}\, [\xi_{i-1}]_{s_i}$, откуда, согласно упорядочению возможностей применения подстановок /см. /3.21/, гл. III/,

/5.27/ $s_i = u^n_{r_{i-1}}(\xi_{i-1}) > u$

или $s_i = u$ и $k_i = \ell^n_{r_{i-1}}(\xi_{i-1}) > k$.

Таким образом

/5.28/ $F: \xi_{i-1} = \rho\, \alpha_k\, \gamma \vdash \xi = F_{ki}^{s_i}(\rho\, \alpha_k\, \gamma) \vdash \eta \in T$,

где, в первом случае $s_i > u = |\rho| + |\alpha_k|$, следовательно слово α_k оказывается продолжительным порядка $|\alpha_k|$ относительно множества T, чем нарушается условие А/ теоремы 9.

Во втором случае для $\ell = k_i > k$ имеет место

/5.29/ $[\xi_{i-1}]^{r_i}\, \alpha_\ell = [\xi_{i-1}]^u = \rho\, \alpha_k$

и, так либо $\alpha_k = [\alpha_\ell]_p$, при $p = |\rho| - r_i \ge 0$

либо $\alpha_\ell = [\alpha_k]_p$, при $p = r_i - |\rho| \ge 0$

Из этого следует нарушение в теореме 9 условия Б/ или В/, соответственно, так как после выполнения соответствующей подстановки получается

/5.30/ $\xi_i = [\rho]^{r_i}\, \beta_\ell\, \gamma$ или $\xi_i = \rho\, [\alpha_k]^p\, \beta_\ell\, \gamma$

и так вывод $F: \xi_i \vdash \eta \in T$ нарушает непродолжимость слов, требуемую в соответствующих условиях.

Таким образом, выполнение А/, Б/ и В/ действительно оказывается необходимым и достаточным для финальности в сильном смысле подстановки $(\alpha_k \to \beta_k)$. Следовательно, они составляют достаточное условие финальности этой подстановки в обыкновенном смысле.

7. Пример. Пусть рассматривается система

$$F=\left\{\begin{array}{lll}(\alpha_1 & = & uxa \to u=\beta_1),\\ (\alpha_2 & = & vyu \to v=\beta_2),\\ (\alpha_3 & = & a \to u=\beta_3),\\ (\alpha_4 & = & u \to v=\beta_4)\end{array}\right\}$$

в алфавите $A=\{a, x, y, u, v\}$.

Множество T состоит из одного элемента v. /Эта система соответствует синтаксису простых арифметических выражений языка АЛГОЛ-60, если

a = ‹множитель›,
u = ‹терм›,
v = ‹простое арифметическое выражение›,
x = ‹знак операции типа умножения›,
y = ‹знак операции типа сложения›,

и для простоты примера выражения типа ± ‹терм› не имеются в виду/.

Последовательный вывод

/5.31/

$$\begin{aligned} F:\ & \xi_0 = \underline{a}\,yaxa \vdash\\ \vdash\ & \xi_1 = \underline{u}\,yaxa \vdash\\ \vdash\ & \xi_2 = v\,y\,\underline{a}\,xa \vdash\\ \vdash\ & \xi_3 = v\,y\,\underline{uxa} \vdash\\ \vdash\ & \xi_4 = \underline{vyu} \vdash\\ \vdash\ & \xi_5 = v \in T \end{aligned}$$

не будет прямым относительно подстановок $(vyu \to v)$ и $(u \to v)$ так как в четвертом шаге вывода, где $r_3 = 2$ имеет место

$$\begin{array}{lll} l_2^1(\xi_3)=2 & U_2^1(\xi_3)=3 & (vyu \to v)\\ l_2^2(\xi_3)=4 & U_2^2(\xi_3)=3 & (u \to v)\\ l_2^3(\xi_3)=1 & U_2^3(\xi_3)=5 & (uxa \to u)\\ l_2^4(\xi_3)=3 & U_2^4(\xi_3)=5 & (a \to u) \end{array}$$

В данном выводе $K_4 = l_2^3(\xi_3) = 1$, т. е применяется третья из этих возможностей, и так, действительно, вывод не будет прямым относительно подстановкам соответствующим первым двум возможностям.

Далее, из этого следует, что эти подстановки не могут быть финальными в сильном смысле относительно данного множества T. Поскольку вывод /5.31/ является единственным выводом, ведущим из слова ауаха в множество T, то эти подстановки, следовательно, будут нефинальными и в обыкновенном смысле.

Покажем, что для подстановки $(vyu \to v)$ действительно нарушается условие А/ теоремы 9, т.е. имеет место $C_F(vyu, 3, T)$. В этом можно убедиться с помощью последовательности типа /5.20/:

/5.32/ $\eta_0 = v,\ d_0 = 0,\ \eta_1 = vyu,\ d_1 = 2,\ \eta_2 = vyu,\ d_2 = 3,$

где $C_F(\eta_0, d_0, T)$ имеет место по условию 5/ определения отношения $C_F(\eta_0, d_0, T)$, а из $C_F(\eta_i, d_i, T)$ следует $C_F(\eta_{i+1}, d_{i+1}, T)$ для $i = 0$ по условию 2/, так как $[vyu]^3 = [\alpha_2]_0$, при $3 = q > r = 2$ и $C_F(\beta_2[vyu]_3, 0, T) = C_F(v, 0, T)$.

Для $i = 1$ это получается с помощью условия 3/, так как

$[vyu]_2 = [\alpha_1]^1 = u$, при $2 = \rho > 0,\ 1 = q < 3 = |\alpha_1|$,

и $C_F([vyu]^2\beta_1, \rho, T) = C_F(vyu, 2, T)$.

Для данной системы F легко можно построить такую систему G, каждая подстановка которой является финальной /даже в сильном смысле/ относительно некоторого множества T' и каждому выводу в системе F, ведущему в это множество, соответствует прямой вывод такой же длины в системе G. Для этого вводится новая буква e, которая пишется в конец каждого слова §, и задается система

$$G=\left\{\begin{array}{ll}(\alpha_1 = axa & \rightarrow u = \beta_1),\\ (\alpha_2 = vyuy & \rightarrow vy = \beta_2),\\ (\alpha_3 = vyue & \rightarrow ve = \beta_3),\\ (\alpha_4 = a & \rightarrow u = \beta_4),\\ (\alpha_5 = uy & \rightarrow vy = \beta_5),\\ (\alpha_6 = ue & \rightarrow ve = \beta_6)\end{array}\right\}$$

и множество $T' = \{ve\}$.

Вывод

/5.33/
$$\begin{array}{l} G: \S_0 = \underline{ay}axae \vdash\\ \vdash \S_1 = \underline{uy}axae \vdash\\ \vdash \S_2 = vy\underline{a}xae \vdash\\ \vdash \S_3 = vy\underline{ux}ae \vdash\\ \vdash \S_4 = \underline{vyue} \vdash\\ \vdash \S_5 = ve \in T'\end{array}$$

соответствующий выводу /5.31/ в системе F, уже будет прямым относительно всех подстановок, так как легко видеть, что во всех шагах выполняется первая возможность применения подстановки.

В финальности подстановок системы G относительно множества T' можно убедиться с помощью теоремы 9. Так, например, для подстановки $(vy_3 \rightarrow vy)$, доказательство отношения $C_G(vyuy, 4, T')$ требовало бы, согласно условию 3/, доказательство отношения $C_G(vyvy, 2, T')$, что - опять с применением того же условия - требует доказательство или самого себя /для $\ell = 2$/, или отношения $C_G(vyve, 2, T')$ /для $\ell = 3$/, а на это последнее, уже ни одно из условий не применимо. Таким образом слова

/5.34/ $\qquad vyuy \qquad\qquad vyvy \qquad\qquad vyve$

и числа

/5.35/ $\qquad 4, \qquad\qquad 2, \qquad\qquad 2$

действительно образуют последовательности типа /5.23/, /5.24/, и так, согласно сказанному, в конце п. 5 ни одно из упомянутых отношений не может иметь место, что обеспечит выполнение условия A/ теоремы 9.

Условие B/ этой теоремы выполняется тривиально, так как ни одно из слов α_ℓ не содержит слово $\alpha_k = vyuy$ а для условия B/ получается

/5.36/ $\alpha_5 = [vyuy]_2$, при $5 = \ell > k = 2$

и так следует доказать $\neg C_G(vyvy, 2, T)$, но это является вторым из отношений, доказанных с помощью /5.34/ и /5.35/

Подобным образом можно показать финальность других подстановок системы G тоже.

VI. ПРИМЕНЕНИЕ МЕТОДА СИНТАКСИЧЕСКОГО АНАЛИЗА В ТРАНСЛЯТОРЕ С ЯЗЫКА АЛГОЛ-60

1. Синтаксическое описание языка АЛГОЛ-60 /см. Наур [12]/, заданное в форме металингвистических формул, называемых иногда "нормальной формой Бэкуса", можно считать примером системы формальных подстановок. При этом металингвистическим формулам типа

/6.1/ $\langle A \rangle ::= \langle B \rangle\langle C \rangle \mid \langle D \rangle$

соответствуют подстановки

/6.2/ $(\langle A \rangle \longrightarrow \langle B \rangle\langle C \rangle)$
$(\langle A \rangle \longrightarrow \langle D \rangle)$

Алфавит системы состоит из совокупности основных символов и металингвистических переменных языка.

Для транслятора, работающего по "синтаксически-управляемому" принципу, требуется проанализировать текст переводимой программы с точки зрения дуальной системы. Цель анализа заключается в том, чтобы найти вывод металингвистической переменной ‹программа› из последовательности основных символов, составляющих переводимую программу. При этом, после выполнения некоторых финальных подстановок, требуется переход к семантическим подпрограммам, которые выполняют задачу самого перевода.

Система формальных подстановок, составленная строго по "официальному" синтаксису языка АЛГОЛ-60, в принципе полностью соответствует этой цели /после исключения всех вхождений металингвистической переменной ‹пустое› /. Но для практической реализации в трансляторе эта система оказывается слишком громоздкой, а применение изложенного метода анализа - в первую очередь из-за нефинальности большого количества подстановок - может потребовать очень много времени. Поэтому ниже дается система формальных подстановок для анализа языка АЛГОЛ-60, имеющая следующие особенности:

а/ при анализе имеется в виду не только чисто синтаксическая структура текста, но также и семантический смысл идентификаторов, заданный с помощью описаний;

б/ подстановки системы являются финальными относительно цели анализа. Для этого используется особенность а/ и требуется полная и правильная спецификация формальных параметров процедур;

в/ подстановки системы задаются в сокращенной форме с помощью таблиц типа /4.16/. При этом условимся расширить обозначения, введенные в гл. IV, следующим образом: если на правой стороне подстановки должны стоять соответствующие элементы некоторых столбцов из таблицы левой стороны, то на правой стороне пишутся соответствующие столбцы таблицы. Например, сокращенная запись

/6.3 $\left\| \begin{matrix} a & & d \\ b & c & e \end{matrix} \right\| \longrightarrow \left\| \begin{matrix} a \\ b \end{matrix} \right\| f$

соответствует подстановкам

/6.4/ $(acd \rightarrow af)$
$(ace \rightarrow af)$
$(bcd \rightarrow bf)$
$(bce \rightarrow bf)$

2. Металингвистические переменные, принимающие участие в система, отличаются от тех, которые используются в официальном описании языка АЛГОЛ-60. Этот факт объясняется тем, что при составлении системы формальных подстановок для практического анализа цель заключается не столько в том, чтобы металингвистические переменные соответствовали некоторым интуитивным понятиям языка, а скорее в том, чтобы их количество было минимальным и финальность подстановок была обеспечена. Поэтому и обозначения металингвистических переменных отличаются от принятых в описании языка АЛГОЛ-60: для обозначения используется сокращение от английского названия того понятия, которое по смыслу стоит близко к данной металингвистической переменной. Приблизительный русский перевод названий этих понятий дается в списке металингвистических переменных.

Кроме металингвистических переменных, заключенных в угловые скобки "<" и ">", в таблицах используются и объекты, заключенные в фигурные скобки "{" и "}". Они не обладают самостоятельным синтаксическим смыслом, а являются только сокращениями списка некоторых элементов алфавита системы /т.е. основных символов или металингвистических переменных/. Название такого сокращения в большинстве случаев совпадает с названием одного из металингвистических переменных данного списка.

Определения этих сокращений даются вместе с подстановками, Так, например, формула

/6.5/ {sae} = <sae> | <term> | <fac> | {apr}

означает, что везде, где в некотором столбце таблицы находится символ {sae}, следует понимать, что в данном столбце стоят металингвистические переменные <sae>, <term>, <fac>, а также все элементы, которые заранее уже были сокращены символом {apr}.

3. Семантический смысл идентификаторов языка АЛГОЛ-60 определяется описаниями. Поэтому, для того, чтобы в анализе вместо понятия "идентификатор" могли принимать участие различные типы идентификаторов, необходима предварительная обработка описаний. Эта обработка выполняется в форме отдельного, самостоятельного просмотра текста переводимой программы, так как в языке АЛГОЛ-60 некоторые идентификаторы могут быть описаны позже, чем их первое вхождение: например, метки /если ее описанием считать вхождение идентификатора перед оператором, разделенное от него символом "двоеточие"/ или нелокальные идентификаторы в процедурах, описанные в том же заголовке блока что и сама процедура.

Для обработки описаний можно также применить метод синтаксического анализа: составляется система формальных подстановок для анализа описаний, и семантические подпрограммы, вызванные в течение анализа при нахождении определенных частей описаний, выполняют все операции, необходимые для составления словаря идентификаторов, имея в виду области действия различных описаний, определенные блочной структурой.

Таким образом получаются две системы: одна для первого просмотра и другая для второго. Первый просмотр должен обрабатывать те части описаний, которые необходимы для определения синтаксического смысла идентификаторов. Очевидно, что некоторые части описаний, как, тело процедуры, переключательный список и граничные пары,могут обрабатываться только во втором просмотре. Таким образом, при выполнении первого просмотра все части программы, не имеющие отношения к обрабатываемым частям описаний, должны оставаться неизменными для обработки во втором просмотре.

Для этого алгоритм анализа в первом просмотре модифицируется следующим образом: при $\varphi = \underline{0}$ все символы, записанные до этого в массив S, передаются на выходную последовательность, которая в свою очередь будет входной последовательностью второго просмотра. /Поскольку подстановки в обоих просмотрах являются финальными, необходимости "возвращения из тупика" в случае $\varphi = \underline{0}$ не может возникнуть/.

Кроме того, на выходную последовательность передаются металингвистические переменные, подставленные вместо частей описаний, обработанных в первом просмотре.

4. При выполнении первого просмотра предполагается, что распознавание основных символов языка АЛГОЛ-60, а также идентификаторов, числовых констант, строк и ограничителей параметров /отличных от запятой/ уже сделано, и вместо этих объектов уже стоят их условные обозначения.

Условные обозначения основных символов и металингвистических переменных состоят из синтаксической и семантической части. Синтаксическая часть в обоих случаях состоит из порядкового номера соответствующего элемента алфавита A системы формальных подстановок. Семантическая часть в случае основных символов покажет, какой имеется в виду из синтаксически равносильных символов, имеющих ту же самую синтаксическую часть /например, x, /, или ↑/, а в случае металингвистических переменных содержит адрес той ячейки памяти, где расположена дальнейшая информация о данном понятии /например, тип и истинный адрес идентификатора, числовое значение константы или предварительный перевод программы вычисления арифметического выражения/.

При некоторых подстановках первого просмотра семантические подпрограммы должны в некотором смысле повлиять на синтаксическое выполнение подстановки. В этих случаях даются соответствующие указания в примечаниях к списку подстановок.

5. Семантические подпрограммы первого просмотра составляют словарь идентификаторов, имеющий в виду блочную структуру программы. При этом в семантическую часть условного обозначения каждого идентификатора пишется адрес некоторой информационной ячейки, содержащей тип и адрес /истинный или относительный/ данного идентификатора, вместе со всей информацией, получаемой от описаний. Таким образом во втором просмотре словарь идентификаторов, содержащий символическую запись идентификаторов, уже не существует, а при выполнении подстановки

$$\| \langle id \rangle \| \rightarrow \{id\}$$

семантическая подпрограмма определяет тип идентификатора, согласно содержанию информационной ячейки.

В системе используются следующие типы идентификаторов:

идентификатор арифметической скалярной величины
" булевской " "
" арифметического массива
" булевского массива
" процедуры
" арифметической функции
" булевской функции
" переключателя
" метки.

Здесь слово "арифметический" одинаково используется в смысле real и integer.

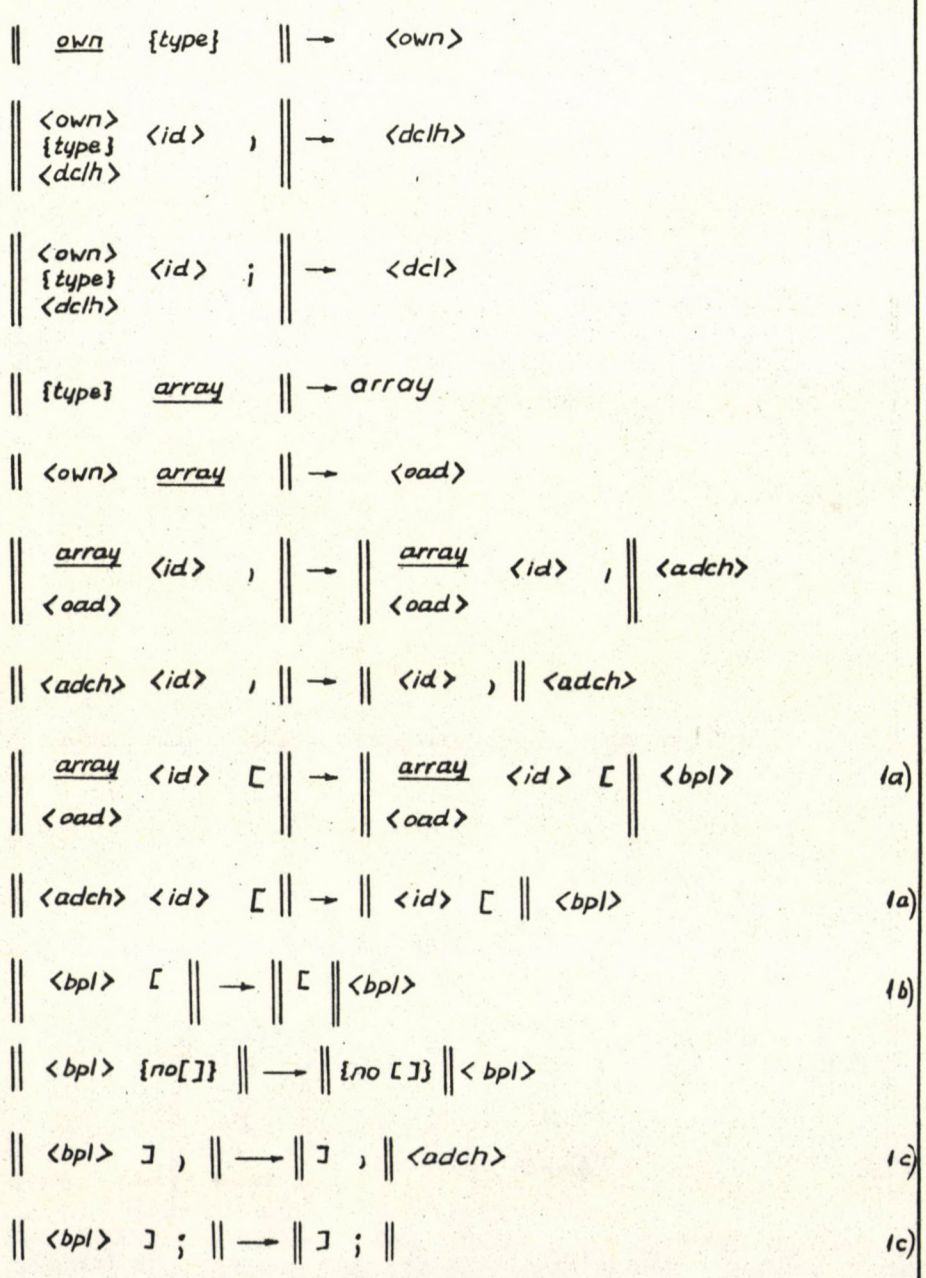
Описанием идентификаторов, играющих роль формальных параметров процедур, считается их спецификация. Поэтому, без требования полной спецификации формальных параметров, тип некоторых идентификаторов оказался бы неопределенным. Для того, чтобы финальность подстановок второго просмотра была и в этом случае обеспечена, требовалось бы значительное усложнение всей системы. Этим объясняется наше требование полной спецификации формальных параметров процедур. При этом, конечно, подразумевается, что спецификация формальных параметров является не только полной, но и правильной в том смысле, что она соответствует той роли, которую данный идентификатор формального параметра играет в теле процедуры.

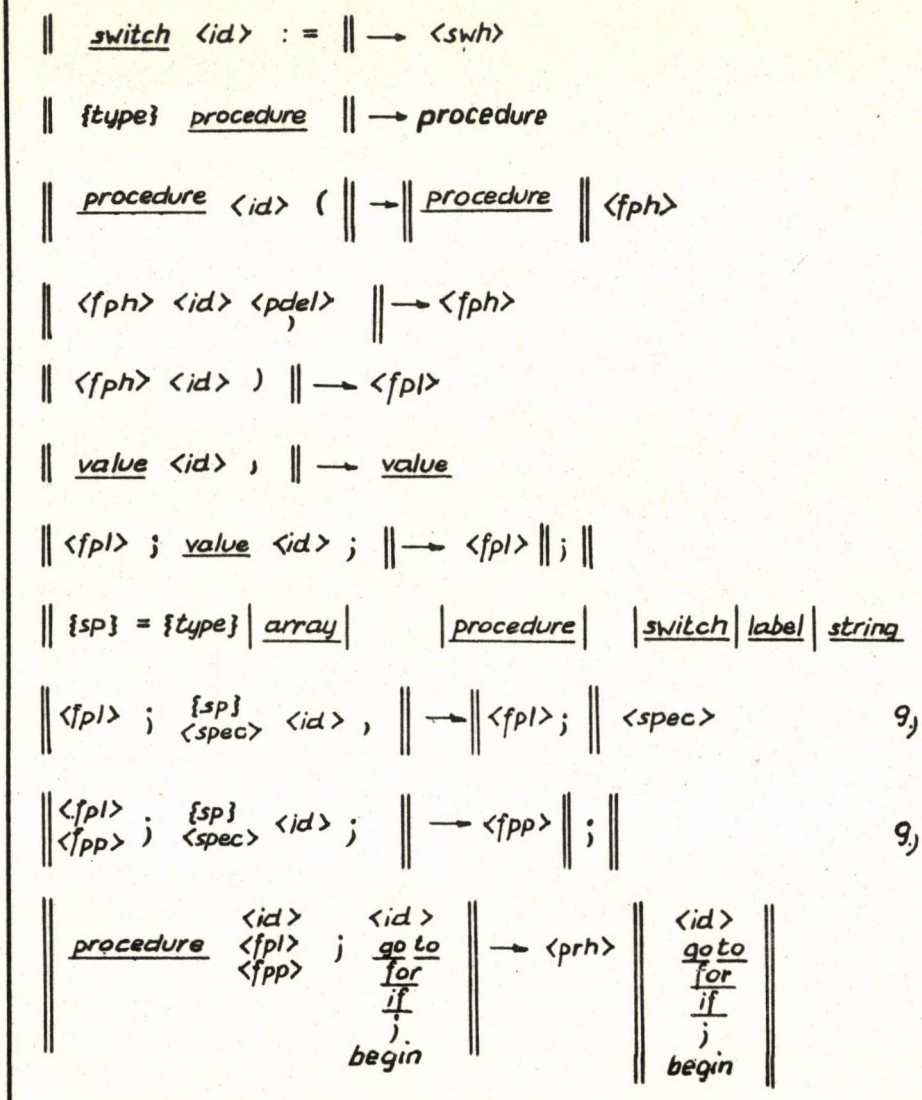
6. Система формальных подстановок для второго просмотра составлена таким образом, что, в случае надобности, второй просмотр можно разбить на три независимые части. В просмотре П.1. обрабатываются выражения и все понятия, которые нужны для распознавания выражений. Просмотр П.2. занимается с понятиями, составленными из выражений, но не использующими понятие оператора; а просмотр П.3. распознает все сложные конструкции из операторов и описаний.

Если второй просмотр организован из таких трех частей, то, конечно, программы просмотров П.1. и П.2. должны быть составлены так, чтобы все символы, необрабатываемые данным просмотром, были переданы следующему просмотру в неизменной форме, т.е. подобно программе первого просмотра.

В практической реализации второго просмотра может оказаться более экономичным объединение в одну подстановку нескольких подстановок, таблицы которых различаются друг от друга только на один символ. В этом случае, отдельные подстановки различаются с помощью семантических подпрограмм. Некоторые такие возможности отмечаются в примечаниях к соответствующим подстановкам.

1.


own {type} → ⟨own⟩
⟨own⟩ {type} ⟨dclh⟩ ⟨id⟩ , → ⟨dclh⟩
⟨own⟩ {type} ⟨dclh⟩ ⟨id⟩ ; → ⟨dcl⟩
{type} array → array
⟨own⟩ array → ⟨oad⟩
array ⟨oad⟩ ⟨id⟩ , → array ⟨oad⟩ ⟨id⟩ , ⟨adch⟩
⟨adch⟩ ⟨id⟩ , → ⟨id⟩ , ⟨adch⟩
array ⟨oad⟩ ⟨id⟩ [→ array ⟨oad⟩ ⟨id⟩ [⟨bpl⟩ 1a)
⟨adch⟩ ⟨id⟩ [→ ⟨id⟩ [⟨bpl⟩ 1a)
⟨bpl⟩ [→ [⟨bpl⟩ 1b)
⟨bpl⟩ {no[]} → {no [ ]} ⟨bpl⟩
⟨bpl⟩] , →] , ⟨adch⟩ 1c)
⟨bpl⟩] ; →] ; 1c)


switch <id> : = → <swh>
{type} procedure → procedure
procedure <id> (→ procedure <fph>
<fph> <id> <pdel> → <fph>
<fph> <id>) → <fpl>
value <id> , → value
<fpl> ; value <id> ; → <fpl> ;
{sp} = {type} | array | | procedure | | switch | label | string
<fpl> ; {sp} <spec> <id> , → <fpl> ; <spec>
<fpl> <fpp> ; {sp} <spec> <id> ; → <fpp> ;
procedure <id> <fpl> <fpp> ; <id> go to for if (begin → <prh> <id> go to for if (begin

{stmbeg} = : | ; | begin | then | else | do | <label> | <prh> | < dcl >

{stmend} = ; | end | else

‖ {stmbeg} {stmend} ‖ ⟶ ‖ {stmbeg} ‖ <bstm> ‖ {stmend} ‖

‖ {stmbeg} <id> : ‖ ⟶ ‖ {stmbeg} ‖ <label>

II. 1.

‖ <id> ‖ ⟶ {id}

{id} = <aid> | < bid > | <aaid> | <baid> | <prid> | <afid> | <bfid> |
<swid> | <lbid>

‖ [<sl> {aexpr} , ‖ ⟶ <sl>

‖ (<apl> {expr} / {id} / <string> <pdel> ‖ ⟶ <apl>

‖ <aaid> [<sl> {aexpr}] ‖ ⟶ <asv> 3)

‖ <baid> [<sl> {aexpr}] ‖ ⟶ <bsv> 3)

‖ <prid> (<apl> {expr} / {id} / <string>) ‖ ⟶ <bstm> 4.)

‖ <afid> (<apl> {expr} / {id} / <string>) ‖ ⟶ <afc> 4.)

$$\| \langle bfid \rangle \; \langle apl \rangle \; \begin{matrix} \{expr\} \\ \{id\} \\ \langle string \rangle \end{matrix} \; \| \longrightarrow \langle bfc \rangle \qquad 4.)$$

$$\{bexe\} =) \mid , \mid \langle pdel \rangle \mid ; \mid \underline{end} \mid \underline{else} \mid \underline{then} \mid \underline{do}$$

$$\{aexe\} = \{bexe\} \mid \wedge \mid \vee \mid \supset \mid \equiv \mid \{rel\} \mid \underline{while} \mid \underline{step} \mid \underline{until} \mid] \mid :$$

$$\| (\{aexpr\}) \| \longrightarrow \langle apr \rangle$$

$$\{apr\} = \langle apr \rangle \mid \langle aid \rangle \mid \langle asv \mid \langle afc \rangle \mid \langle afid \rangle$$

$$\left\| \begin{matrix} \{apr\} \\ \langle fac \rangle \uparrow \end{matrix} \; \{apr\} \; \begin{matrix} \{aexe\} \\ \{add\} \\ \{mult\} \uparrow \end{matrix} \right\| \longrightarrow \langle fac \rangle \left\| \begin{matrix} \{aexe\} \\ \{add\} \\ \{mult\} \uparrow \end{matrix} \right\|$$

$$\left\| \begin{matrix} \{apr\} \\ \langle fac \rangle \\ \langle term \rangle \end{matrix} \; \{mult\} \; \begin{matrix} \{apr\} \\ \langle fac \rangle \end{matrix} \; \begin{matrix} \{aexe\} \\ \{add\} \\ \{mult\} \end{matrix} \right\| \longrightarrow \langle term \rangle \left\| \begin{matrix} \{aexe\} \\ \{add\} \\ \{mult\} \end{matrix} \right\|$$

$$\left\| \{sae\} \; \{add\} \; \begin{matrix} \{apr\} \\ \langle fac \rangle \\ \langle term \rangle \end{matrix} \; \begin{matrix} \{aexe\} \\ \{add\} \end{matrix} \right\| \longrightarrow \langle sae \rangle \left\| \begin{matrix} \{aexe\} \\ \{add\} \end{matrix} \right\| \qquad 5.)$$

$$\left\| \{add\} \; \begin{matrix} \{apr\} \\ \langle fac \rangle \\ \langle term \rangle \end{matrix} \; \begin{matrix} \{aexe\} \\ \{add\} \end{matrix} \right\| \longrightarrow \langle sae \rangle \left\| \begin{matrix} \{aexe\} \\ \{add\} \end{matrix} \right\| \qquad 5.)$$

$$\{sae\} = \langle sae \rangle \mid \langle term \rangle \mid \langle fac \rangle \mid \{apr\}$$

$$\| \langle ifcl \rangle \; \{sae\} \; \underline{else} \; \{aexpr\} \; \{aexe\} \| \longrightarrow \langle aexpr \rangle \| \{aexe\} \|$$

$$\{aexpr\} = \langle aexpr \rangle \mid \{sae\}$$

$$\| (\{bexpr\}) \| \longrightarrow \langle bpr \rangle$$

$\| \{sae\}\ \{rel\}\ \{sae\}\ \{aexe\} \| \rightarrow \langle bpr \rangle \| \{aexe\} \|$

$\{bpr\} = \langle bid \rangle \mid \langle bsv \rangle \mid \langle bfc \rangle \mid \langle bfid \rangle \mid \langle bpr \rangle$

$\| \neg \{bpr\} \begin{matrix} \{bexe\} \\ \wedge \\ \vee \\ \supset \\ \equiv \end{matrix} \| \rightarrow \langle bsec \rangle \| \begin{matrix} \{bexe\} \\ \wedge \\ \vee \\ \supset \\ \equiv \end{matrix} \|$

$\| \begin{matrix} \{bpr\} \\ \langle bsec \rangle \\ \langle sbe1 \rangle \end{matrix} \wedge \begin{matrix} \{bpr\} \\ \langle bsec \rangle \end{matrix} \begin{matrix} \{bexe\} \\ \wedge \\ \vee \\ \supset \\ \equiv \end{matrix} \| \rightarrow \langle sbe1 \rangle \| \begin{matrix} \{bexe\} \\ \wedge \\ \vee \\ \supset \\ \equiv \end{matrix} \|$ 6.)

$\| \begin{matrix} \{bpr\} \\ \langle bsec \rangle \\ \langle sbe1 \rangle \\ \langle sbe2 \rangle \end{matrix} \vee \begin{matrix} \{bpr\} \\ \langle bsec \rangle \\ \langle sbe1 \rangle \end{matrix} \begin{matrix} \{bexe\} \\ \vee \\ \supset \\ \equiv \end{matrix} \| \rightarrow \langle sbe2 \rangle \| \begin{matrix} \{bexe\} \\ \vee \\ \supset \\ \equiv \end{matrix} \|$ 6.)

$\| \begin{matrix} \{bpr\} \\ \langle bsec \rangle \\ \langle sbe1 \rangle \\ \langle sbe2 \rangle \\ \langle sbe3 \rangle \end{matrix} \supset \begin{matrix} \{bpr\} \\ \langle bsec \rangle \\ \langle sbe1 \rangle \\ \langle sbe2 \rangle \end{matrix} \begin{matrix} \{bexe\} \\ \supset \\ \equiv \end{matrix} \| \rightarrow \langle sbe3 \rangle \| \begin{matrix} \{bexe\} \\ \supset \\ \equiv \end{matrix} \|$ 6.)

$\| \{sbe\} \equiv \begin{matrix} \{bpr\} \\ \langle bsec \rangle \\ \langle sbe1 \rangle \\ \langle sbe2 \rangle \\ \langle sbe3 \rangle \end{matrix} \begin{matrix} \{bexe\} \\ \equiv \end{matrix} \| \rightarrow \langle sbe \rangle \| \begin{matrix} \{bexe\} \\ \equiv \end{matrix} \|$ 6.)

$\{sbe\} = \{bpr\} \mid \langle bsec \rangle \mid \langle sbe1 \rangle \mid \langle sbe2 \rangle \mid \langle sbe3 \rangle \mid \langle sbe \rangle$

$\| \langle ifcl \rangle\ \{sbe\}\ \underline{else}\ \{bexpr\}\ \{bexe\} \| \rightarrow \langle bexpr \rangle \| \{bexe\} \|$

$\{bexpr\} = \langle bexpr \rangle \mid \{sbe\}$

‖ *if* {bexpr} *then* ‖ → ⟨ifcl⟩

‖ ⟨swid⟩ [{aexpr}] ‖ → ⟨sdes⟩

‖ ({dexpr}) ‖ → ⟨sdes⟩

‖ ⟨ifcl⟩ ⟨lbid⟩ ⟨sdes⟩ *else* {dexpr} ‖ → ⟨dexpr⟩

{dexpr} = ⟨dexpr⟩ | ⟨sdes⟩ | ⟨lbid⟩

{expr} = ⟨aexpr⟩ | ⟨bexpr⟩ | ⟨dexpr⟩

II. 2.

‖ ⟨aid⟩ ⟨asv⟩ ⟨afid⟩ := {aexpr} {stmend} ‖ → ⟨aass⟩ ‖ {stmend} ‖

‖ ⟨aid⟩ ⟨asv⟩ ⟨afid⟩ := ⟨aass⟩ ‖ → ⟨aass⟩

‖ ⟨bid⟩ ⟨bsv⟩ ⟨bfid⟩ := {bexpr} {stmend} ‖ → ⟨bass⟩ ‖ {stmend} ‖

‖ ⟨bid⟩ ⟨bsv⟩ ⟨bfid⟩ := ⟨bass⟩ ‖ → ⟨bass⟩

‖ *go to* {dexpr} ‖ → ⟨bstm⟩

{bstm} = ⟨bstm⟩ | ⟨aass⟩ | ⟨bass⟩ | ⟨prid⟩ | ⟨afid⟩ | ⟨bfid⟩

⟨afc⟩ | ⟨bfc⟩

| {aexpr} step {aexpr} until {aexpr} do , | → <forel> | do ⌉ |

| {aexpr} while {bexpr} do , | → <forel> | do , |

| for <aid> <asv> : = | → <for>

| <for> <forel> <aexpr> , | → <for>

| <swh> {dexpr} , | → <swh>

| <swh> {dexpr} ; | → <dcl>

| [<bph> {aexpr} : {aexpr} , | → <bph>

| <aaid> <baid> [<bph> {aexpr} : {aexpr}] | → <adel>

| array <ad> <oad> <aaid> <baid> <adel> , | → <ad>

| array <ad> <oad> <adel> ; | → <dcl>

II. 3.

| <for> <forel> {aexpr} do {stm} | → <forstm>

| <ifcl> {bstm} | → <ifstm>

| <ifstm> else {stm} | → <cstm>

‖ <ifcl> <forstm> ‖ → <cstm>

‖ begin <dcl> ‖ → begin

‖ begin / <bcsh> {stm} ; ‖ → <bcsh>

‖ begin / <bcsh> {stm} end ‖ → <bstm>

{stm} = <bstm> | <forstm> | <ifstm> | < cstm>

‖ <prh> {stm} ; ‖ → <dcl>

‖ <label> {stm} ‖ → ‖ {stm} ‖ 7.)

‖ {stm} ; ‖ → <progr> 8.)

СИСТЕМА ФОРМАЛЬНЫХ ПОДСТАНОВОК ДЛЯ АНАЛИЗА ЯЗЫКА АЛГОЛ-60

1. Алфавит системы

1.1. Основные символы языка, группированные согласно их синтаксическому значению:

```
add    = +    -     /знак операции типа сложения/
mult   = x    /     ÷   /знак операции типа умножения/
rel    = <  ≤  =  ≥  >  ≠        /знак отношения/
```

```
bpr    = true  false      /булевское первичное выражение/
         if
         go to
         for
         until
         step
         while
         do
         ther
         begin
         else
         end
```

```
         :
         :=
         ;
         ,
type   = real   integer   Boolean
                /описатель типа/
         array
         switch
         procedure
          spec    = label   string
         value
         own
```

1.2. Составные конструкции распознанные вместе с основными символами еще до первого просмотра /см. п. 4, гл. VI/:

‹id›	идентификатор
‹apr›	первичное арифметическое выражение /т.е. в этом случае константа/
‹string›	строка
‹pdel›	ограничитель параметра

1.3. Металингвистические переменные первого просмотра.

1.3.1 Переменные, передаваемые второму просмотру

‹dcl›	декларация
‹oad›	описатель собственного массива
‹bstm›	основной оператор
‹label›	метка
‹swh›	заголовок описания переключателя
‹prh›	заголовок описания процедуры

1.3.2 Остальные переменные первого просмотра:

‹own›	собственный тип
‹dclh›	начало описания типа
‹adch›	начало описания массива
‹bpl›	список граничных пар
‹fph›	начало списка формальных параметров
‹fpl›	список формальных параметров
‹fpp›	формальные параметры и их спецификации

1.4 Металингвистические переменные второго просмотра /вместе с символами 1.1 и 1.2 и переменными 1.3.1/:

1.4.1 Семантические типы идентификаторов:

‹aid›	арифметический скаляр
‹bid›	булевский скаляр
‹aaid›	арифметический массив
‹baid›	булевский массив
‹prid›	процедура
‹afid›	арифметическая функция
‹bfid›	булевская функция
‹swid›	переключатель
‹lbid›	метка

Вместо идентификаторов, специфицированных как string подставится переменная ‹string› , так как эти идентификаторы могут стоять в программе только на таких местах как строки.

‹sl›	начало списка индексов
‹apl›	начало списка фактических параметров
‹asv›	арифметическая переменная с индексом
‹bsv›	булевская переменная с индексом
‹afc›	арифметическая функция
‹bfc›	булевская функция
‹fac›	множитель
‹term›	терм
‹sae›	простое арифметическое выражение
‹aexpr›	арифметическое выражение
‹bsec›	вторичное булевское выражение
‹sbe 1›	булевский множитель
‹sbe 2›	булевский терм

‹sbe 3›	импликация
‹sbe›	простое булевское выражение
‹bexpr›	булевское выражение
‹ifc|›	условие
‹sdes›	простое именующее выражение
‹dexpr›	именующее выражение

1.4.3 Переменные, дополнительно введенные просмотром П.2:

‹aass›	арифметический оператор присваивания
‹bass›	булевский оператор присваивания
‹forel›	элемент списка цикла
‹for›	заголовок цикла
‹bph›	начало списка граничных пар
‹adch›	начало описания массива
‹adel›	сегмент массива

1.4.4 Переменные, дополнительно введенные просмотром П.3:

‹forstm›	оператор цикла
‹ifstm›	оператор "если"
‹cstm›	условный оператор
‹bcsh›	начало блока или составного оператор
‹progr›	программа

1.5 Сокращения /название которых не совпадает с названиями одного из металингвистических переменных/:

{stmbeg}	символ, стоящий перед оператором
{stmend}	символ, стоящий после оператора
{aexe}	символ, стоящий после арифметического выражения
{bexe}	символ, стоящий после булевского выражения
{sp}	спецификатор
{expr}	выражение
{stm}	оператор

3. Примечания к подстановкам

1/ Для того, чтобы найти закрывающую квадратную скобку, соответствующую открывающей скобке списка граничных пар, необходимо подсчитывать все квадратные скобки внутри этого списка. Поэтому при выполнении подстановок 1а посылается нуль в некоторую ячейку-счетчик s, а перед выполнением подстановки 1в единица прибавляется или вычитается в зависимости от того, что скобка была открывающей или закрывающей. Сама подстановка 1в выполняется только в случае $s \geq 0$, а если нет, то после следующего символа /, или ;/ обязательно выполняется одна из подстановок 1с.

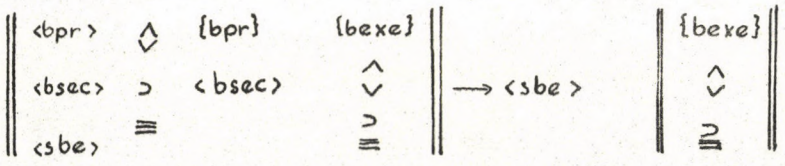
2/ Согласно содержания информационной ячейки, заполненной в первом просмотре, выбирается и подстанавливается соответствующий семантический тип из списка {id}

3/ Можно соединить эти две подстановки в одну, таблица которой содержит в первом столбце символы ⟨aaid⟩ и ⟨baid⟩ а семантическая подпрограмма должна определить, какой из символов ⟨asv⟩ и ⟨bsv⟩ подстанавливается.

4/ Подобно случаю в примечании 3/ можно соединить и эти подстановки, различающиеся только в первом столбце таблицы.

5/ Для того, чтобы обеспечить финальность, первая из этих подстановок должна иметь порядковый номер меньший, чем вторая.

6/ Для сокращения размеров системы вместо этих четырех подстановок можно писать подстановку



а учет порядка старшинства операций поручить семантической подпрограмме.

7/ В результате применения этой подстановки символ метки совсем исчезнет из последовательности символов, и так семантическая подпрограмма должна заботиться о том, чтобы при работе транслятора имелся в виду тот факт, что данная метка стояла на этом месте.

8/ Эта подстановка, выполнение которой означает конец анализа, должна иметь наибольший порядковый номер и так выполняться только в том случае, если ни одна из других подстановок не применима.

9/ Эти подстановки должны иметь порядковый номер меньше чем соответствующие подстановки описаний типа /вторая и третья таблицы на стр.90/, чтобы обеспечить правильную работу в случае спецификаций, имеющих форму описания типа.

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] CHOMSKY, N.: On certain formal properties of grammars. Information and Control, Vol. 2, N° 2, 1959.

[2] BROOKER, R. and MORRIS, D.: Trees and routines. The Computer Journal, Vol. 5, N° 1, 1962. Apr.

[3] ЖОГОЛЕВ, Е.А.: Алгоритм выделения понятий с помощью синтаксической таблицы. Журнал вычислительной математики и математической физики 1965 г. том 5, № 4, стр. 689-697.

[4] KUNO, S. and OETTINGER, A.: Multiple path syntactic analyzer. Information Processing 62, North Holland, Amsterdam, 1963.

[5] GREIBACH, S.: Inverses of phrase structure generators. Mathematical Linguistics and Automatic Translation, Rep. N° NSF-11, Harvard Comp. Lab.

[6] IRONS, E.: A syntax directed compiler for ALGOL-60 Communications of the ACM, Vol. 4, N° 1, 1961.

[7] GRIFFITHS, T.V. and PETRICK, S.R.: On the Relative Efficiencies of Context-Free Grammar Recognizers. Communications of the ACM, Vol. 8, N°5,1965

[8] МАРКОВ, А.А.: Теория алгорифмов. Труды математического института АН СССР, том 42, 1954 г.

[9] PETER, R.: Über die Veralgemeinerung der Theorie der rekursiven Funktionen für abstrakte Mengen geeigneter Struktur als Definitionsbereiche. Acta Math. Acad. Hung. I. Teil: 12 (1961) 271-314, II. Teil: 13 (1962) 1-24.

[10] PETER, R.: Über die Rekursivität der Begriffe der mathematischen Grammatiken. MTA Mat. Kut. Int. Közleményei 8/A (1963) 213-228.

[11] MAZURKIEWICZ, A.: O problemie czytania dla pewnych jezykow formalnych. Instytut Maszyn Matematycznych, Warszawa, 1964. /диссертация, по польски/

[12] NAUR, P. (ed.): Revised Report on the Algorithmic Language ALGOL-60. Communications of the ACM, Vol. 4, N° 1, 1963.

[13] DÖMÖLKI, B.: Jelsorozatok tulajdonságainak felismerésére szolgáló algoritmusok. MTA Számitástechnikai Központ Tájékoztatója, N° 8, 1962. /по венгерски. Краткое содержание, написанное Ф. Кифером см. в Computational Linguistics N° 1, 1963/.

[14] ДОМЕЛКИ, В.: Алгоритмы для распознавания свойств последовательностей символов. Журнал вычислительной математики и математической физики, 1965 г., том 5, № 1.

[15] DÖMÖLKI, B.: An algorithm for syntactic analysis. Computational Linguistics N° 3, 1964.

UTILIZATION OF LEXICAL KNOWLEDGE IN AUTOMATIC TRANSLATION

Gy. Hell

Problems of a semantic nature cropped up already at the very outset of the history of mechanical translation, and just like today, they proved to be a very formidable task [1]. In mechanical translation semantic problems appear in two specific forms, one of them associated with the method of word-by-word translation, the other with translation by sentence analysis.

Semantic problems of the word-by-word translation

In a word-by-word translation owing to the natural multiple meaning of the linguistic elements each word or morpheme of the source language may in general produce several words or morphemes with appropriate meanings in the target language which will not only puzzle the reader, but at the same time frustrate the understanding of the sentence. The method of word-by-word translation tries to improve the quality of the output text by avoiding as far as possible all ambiguities, or by reducing their number at least. Within this method this can be achieved if the purely grammatical and purely semantic ambiguities are separated by considering the morphemes or words translated in isolation from others as grammatical or lexical elements [2]. In this way the grammatical ambiguities can be narrowed down by studying the immediate environment, the semantic ambiguities by keeping in mind the character of the content of the whole text. The improved methods of word-by-word translation rely on this principle when the word of the target language is chosen by an analysis of the grammatical properties of the elements before and after the morpheme under study and by the use of the meaning given in the micro-glossary.

Semantic problems of translation by sentence analysis

It was not on account of the semantic difficulties outlined above that the word-by-word translation proved insufficient. The word-by-word translation from one language to another is only possible if the word orders of the languages do not differ considerably. Whenever the word order of the target language greatly differs from that of the source language and the word order of the former can only be arrived at by an analysis of the syntactical function of the words, the word-by-word method has to be discarded. In this case (e.g. translations from Russian to Hungarian) the object can only be attained by syntactic analysis instead of an analysis of the words or environments of the words of the source language [3].

The application of syntactic analysis has not solved the problems of automatic translation. What's more syntactic analysis has discovered characteristics of the semantic properties of a language, that in earlier analytical work have not turned up. For the speaker and listener it is beyond doubt that apart from specific interjections each word of a sentence becomes a unit only within a uniform system of relationships. In this unit a single dependency line runs from the element on the peak of the system of relationships to the elements that have no further dependent members. There are no closed sections

in the dependency tree of a sentence). In a great number of sentences formal analysis is unable to shape these units without error and none of the methods of syntactical sentence analysis could produce an algorithm which would have yielded a complete sentence analysis in a satisfactory manner.

The picture of the interdependences within the sentence as drawn by the dependency grammar does not ignore the cases in which more than a single path leads from the elements on the peak of the sentence to the bottom-most elements. Such sentences with grammatically multiple meaning must as a matter of course be valued as such of multiple meaning also in formal analysis. However, it is not this type of sentences that throws obstacles in the way of automatic translation. As a matter of fact the difficulties are caused by sentences which are unequivocal for both speaker and listener and whose analysis involves no particular problems for them. Yet when dealing with sentences of this type the automatic translator will put out several closed syntactic structures and not a single one, i.e. the translator will attribute multiple meaning to these sentences.

Multiple meaning obtained by syntactic analysis is different from that occurring in word-by-word translation. At first glance this multiple meaning may appear to be of a grammatical nature, as it applies to the structure of the sentence and it has been discovered in the course of grammatical analysis. However, here the grammatical character presents traits altogether different from those discovered in a word-by-word translation. Here it is not a case of isolated morphemes, and the problem that has to be settled is not whether these morphemes are grammatical, and if so, which of the multiple meaning is the one that on account of the environments has to be considered unique. It is a problem of the mutual relation of free morphemes, which are partly produced on account of the grammatical properties of the elements participating in the relation, but cannot be deduced wholly from these properties by using the earlier methods. These relations are also semantic in character and their solution cannot be expected unless semantics also are introduced in the analysis of automatic translation.

The type of sentences that present the greatest difficulties in automatic translation are

$$N_{at} \qquad V_{tr} \qquad N_{at}$$

where V_{tr} is a transitive verb,

N_{at} is a noun which by formal criteria may equally be a nominative or accusative. Such sentences are e.g. in Russian:

/1/ Билет выдает автомат

/2/ Кислород доставляет кровь

/3/ Радость вызывает признание

/4/ Радость вызывает успокоение

On the ground of the notions of the structural grammars sentences of this type will lend themselves to an analysis on formal grounds with success only if the word categories can be split up into sub-categories by means of which nouns in the nominative and accusative can be distinguished from each other. There is no difficulty in finding such discriminatory sub-categories for the first sentence, however, difficulties arise in the second sentence. In the

third and fourth sentences the word "радость" by the side of the verb "вызывает" is once in the accusative, once in the nominative. It is extremely difficult to find a categorization which in sentences of this type would help us to a satisfactory analysis.

Another type of sentences constituting a serious problem for automatic translation is represented by the following formula:

$$N_a \quad V_z \quad N_x \quad Pr \quad N_y$$

Here V_z is a verb by the side of which there may be nouns $Pr\ N_y$ as well as nouns N_x, however, N_y not as an obligatory government. At the same time this $Pr\ N_y$ may be a dependent member of the noun immediately preceeding it. E.g.

/5/ Я сижу в комнате с моими друзьями

/6/ Я сижу в комнате с большими окнами

In these sentences, besides the complete identity of subject and predicate, there is also a formally uniform construction, which behaves in an entirely different manner in the two sentences. For the speaker both sentences are unequivocal, for the machine they are equivocal. At the very outset of automatic translation Brandwood pointed out the difficulties implied in the formal analysis of such adjectives of adverbial form. However, only papers published of late discuss the potentialities of a grammatical analysis of this phenomenon exhaustively [4].

In formal analysis for the analysis of adjectives of an adverbial form apart from the sub-categories of word categories use is made of the various types of verbal and substantical governments, the rules of word order given by projective dependency, the list and table of idiomatic expressions or set phrases in a context of permanent character and definite subject-matter, consistencies recognisable in the actual pattern of a sentence, and of an ultimate means of formal dependency, i.e. the statictical data of distribution. Transformational analysis which is an important formal means of structural grammar, could not as yet be used in automatic analysis.

Since with the methods of analysis actually available for the purpose of automatic translation not all sentences can be analysed without error, the papers dealing with the analysis of the sentences of multiple meaning here presented are of the opinion that semantic relations cannot be ignored in an attempt to solve the grammatical problem.

Semantic systems in automatic translation

In recent years several attempts were made to build up a semantic system which could be utilized with success in automatic translation. These attempts in general did not go back to theories of semantics already formulated in general linguistics, but by relying mostly on the practice of lexicography set off from groups of synonymous words, or the definitions of meaning of the dictionaries. Apart from the strongly grammatical system of sememes of S. Lamb and P. Sgall there are two parctical systems worked out in detail at present, viz. the system called thesaurus, and the method of definition i.e. defining the meaning of the word by enumerating the elements of meaning [5].

Several objections have been raised calling into doubt the usefulness of a thesaurus [6]. One of the objections is that the groups of the thesaurus

have been formed in an arbitrary manner, based on some subjective straightforwardness, and it is by no means certain that these groups lend themselves to practical uses. This objection is by no means of an accidental one, as automatic translation insists on accurately formulated rules that are based on the properties of the linguistic material, and not on an individual opinion. Research work performed by the Cambridge group tends to confirm that the method of Roget cannot be used without amendments. However, the amendment of a thesaurus is by no means a simple matter. If the basic elements have to be regrouped, this cannot be done with only a part of the elements. As a matter of fact the system once created has to be transformed completely. Several studies are in progress with an end in view of constructing groups of the thesaurus on the basis of textual analysis, in the possession of statistical data [7] . There are no completed and serviceable systems available in the literature of automatic translation as yet.

Two semantic systems of definitions based on the practices developed in the compilation of dictionaries may be distinguished. One system relies on the elements of the definition of meaning resembling the features of the concept, the other utilizes those words of the definition whose subsequent interpretation does not amount to a repetition of words [8].

So far the systems of definition have failed to produce serviceable rules for the purpose of automatic translation. The reason may lie in that though these systems make efforts to be as concrete as possible, still in the formulation of their rules they lack the precision required for writing down satisfactory algorithms. So e.g. one can never see clearly which are the semantic distinguishers by which the meaning of a word may be defined. In the trials made so far a variety of distinguishers have been used [9].

As a matter of fact it is essential that the distinguishers define each word in an adequate manner, however, at the same time it is not allowed to distinguish each word individually, for in this case the system would become impracticable. Hence the limits must be drawn for the distinction of the particular words, and it is exactly this that makes the task difficult.

A specific difficulty lies in the formulation of the connecting rules of words provided with semantic distinguishers. This does not merely imply the difficulty involved in the mathematical coordination of the matrices incorporating the rows of the characteristics of a word (this problem has not as yet been solved in a satisfactory manner [10], but also the difficulty involved in drawing up the locutions or word groups correctly. The studies made so far in this field tried to point at the possibility of a solution only within the narrow sphere of examples selected at random.

The factual character of semantics

If the question has to be answered which of the two systems, i.e.the thesaurus or the system of definition meets the requirements of automatic translation better,it is hard to arrive at a clearcut answer (the answer will in like way be an unsatisfactory one). Whereas for formal consideration, the thesaurus appears the more reassuring, yet it will fail to solve the most difficult problems, i.e. the problems of homonymous syntactic relations in automatic

translation. As a matter of fact synonymous groups do not determine the existence, i.e. possibility or impossibility of locutions, but yield the harmony of words surveyed in isolation. This involves the selection of that meaning of the words which from the point of view of the topic is satisfactory. The synonyms help at the selection of the appropriate lexical unity of the text of the target language, but do not touch the problem of the syntagmatic-syntactic idioms.

It is exactly this latter problem where the semantic systems of definition are rather promissing. However, it is doubtful whether the semantic signs of the words as used in the system are adequate for the purpose. As a matter of fact the semantic signs of the word are not directly given in their entity, but come forth only in the course of a connexion with or in a juxtaposition to other words according to the environment. (This does not necessarily mean that the words receive their meaning only in their context.) It is also essential that among the signs taken as a basis by an interpretative definition there are not only quantitative or qualitative signs, but also spatial, temporal or topical-empirical ties, which consistently come forth in locutions or idioms, but which it would for practical purposes be an exaggerated demand to reproduce in signs. So e.g. the sentence

/7/ Он читает доклад о машинном переводе в Венгрии will admit an interpretation (and a grammatical analysis) in the sense that a lecturer (not in Hungary) speaks of automatic translation in Hungary only if we know that there is automatic translation in Hungary. (There are many sentences of this type.)

It is perhaps a better approach to problems of a semantic nature if the requirement of factual information is construed not to make human thinking (which is supposed to be of a semantic nature) formal, but rather to make topical knowledge factual. At the first glance this approach may appear absurd, for it presupposes the accumulation and storage of a tremendous mass of facts which is unattainable.

In practice this mass of facts will take on a size (whithin graps) that is managable still.

Automatic translation is the applied field of a discipline seeking relations of a more general character and validity. In the course of the solutions of semantic problems this applied character will prove useful, as it allows a search for a practical solution by narrowing down the multitude of phenomena which are otherwise hard to define. As we have seen it is the definition of locutions or idioms that introduces difficulties into automatic translation by sentence analysis. This means that the unities of the topical mass of facts have to be studied in their relations. Of course relation here must necessarily be a relation expressed by the language, i.e. the arrangement of words made possible by the language. It stands to reason that there is an enormous number of such relations, however, automatic translation has no need for all possible relations. When translation between two languages is considered, only those groups of relations will be needed which in the source language have not been expressed by grammatical means. So e.g. (when Russian as a source language is kept in view) it is not necessary to provide the adjective-noun relations.

The subject-predicate-object relation is not always unequivocal at the formal-grammatical analysis of a Russian text. The semantic programme of automatic translation incorporates all the rules which unequivocally yield the connexion of words that are important because of their multiple meanings.

Hence the semantic program does not extend to all nouns, but only to those whose subject-object relation cannot be determined on a grammatical basis. These nouns are the masculine inanimates, the feminines ending in the soft mark, and the neutrals, within the corresponding micro-glossary. These nouns are merely the terms of the subject-object relation, the relation itself is provided by the verbal word. The semantic program compiles the possible relations between the necessary nouns in tables each grouped round a verbal word as a relation carrying element.

Let the following sentences be given, in which the syntactical relations are to be determined from the underlined words:

Пример дает объяснение
Прилежность дает успех
Значение а дает давление I
Толчок дает скорость
Билет дает автомат

With the nouns the following matrix may be formed for the verb "дать" as the basic member of the relation:

	1	2	3	4	5	6	7	8	9	10
1. пример		+	-	-	+	+	-	+	-	-
2. успех	-		-	-	+	+	-	+	-	-
3. прилежность	+	+		+	-	+	-	+	-	-
4. билет	-	-	-		-	+	-	+	-	-
5. деление	-	-	-	-		-	-	+	-	-
6. толчок	-	+	-	-	-		+	+	-	+
7. скорость	-	-	-	-	+	+		+	-	-
8. объяснение	-	+	-	-	-	-	-		-	-
9. автомат	-	+	-	+	-	+	+	-		+
10.давление	-	-	-	-	+	+	+	+	-	

This table provides two-valued relations, however, tables containing multi-valued relations may also be compiled:

two-valued table: A- дать - T

three-valued table: A:T - дать - R

This latter is similar in form of the former table, however, the rows yield the possible A:T relations, whereas the columns contain the nouns possible for purposes of an R-connexion. (For the purpose of an analysis of the Russian text the relations A-B, or A:T - R require no analysis.)

As it has already been seen prepositional nouns, when standing immediately after the noun often form homonymous constructions, and may be members dependent on the verb or on the preceeding noun. Matrices are well suited to

solve problems of this type. In this case the basic relation will be formed by a verb in conjunction with a preposition and not a verbal word by itself.

It is by no means necessary to give the relations in a matrix form. E.g. the А ДАТЬ-Т relation can be expressed also in the following form:

А – ДАТЬ-Т 1; 2,5,6,8
2; 5,6,8
3; 1,2,4,6,8
etc.

The tables so compiled may also be used for the formation of groups of synonyms. In this case the words of a similar semantic interpretation will of course refer to the basic member of the relation, and consequently modify the table, e.g.:

:
.
4,9,12 ; 5,7,8,13,20 ...
.
:

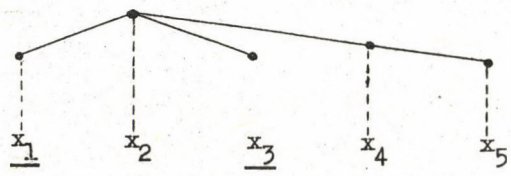
The number of tables so obtained depends on the number of words providing the basis of the relation. On the ground of synonymity even the basic words may form groups, although the majority of them will be unique words.

The relational system of the topical data does not strain the dictionary. The words acting as the terminals of the relations are taken up in the tables of relations with their serial numbers in the dictionary, and only the words of multiple meaning receive a new serial number or numbers by the side of their current serial number.

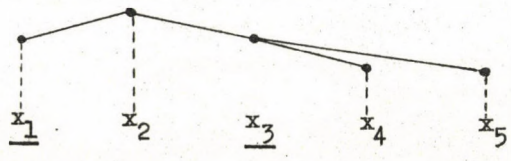
The system of relations provided means for the formal solution of transformational potentialities. Let us consider the following sentence e.g.: Автомат дает билет с объяснением скорости.

With the preposition attached to the noun there are five elements in this sentence, whose potential relations of dependency may be represented as follows:

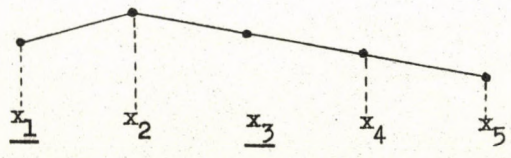
1)



2)



3)



Here owing to the formal identity the relation of x_1 and x_3 to x_2 has not as yet been cleared in all three cases. With the aid of the tables of relations

not only this problem may be solved in the manner already described, but also others.

Fig. 1 is possible when in the Table A - дать - T there were also 9;9 by the side of 9;6 (which in certain topics is perhaps possible).

Fig. 2 is possible when there is 4;8 in the Table A - дать - T, or when on the basis of a Table A - объяснять - T the sentence билет объясняет скорость is possible. These are, however, problems of transformation.

A basis for Fig. 3 of dependency is provided by other transformational potentialities.

The factual arrangement of the topical body of knowledge in the manner here described is a by no means simple matter. It does not even promise to call on the memory of the translator with moderation only. However, this arrangement has properties, that are advantageous:

/1/ it lays the main stress on the relation of the linguistic elements and by this the solution of the focal problems of syntactic analysis may be attempted;

/2/ the tables containing the relations rely on the interpretation of the meaning, which is also the rule of the possible connections of the words;

/3/ within a given dictionary the connection of the words may be given in a way characteristic of a given topic;

/4/ the rules of the connection of the words cannot be ignored even by the formal semantic systems of the definition type, however, these systems attach the rules to the hard to recognice semantic distinguishers of the words among which the spatial and temporal restrictions cannot be included. The system of relations relying on the body of topical knowledge ignores the semantic signs altogether;

/5/ since automatic translation is an applied linguistic task, it may be convenient to make use of potential simplifications incidental to the specific circumstances of application. Simplification is made possible by narrowing down the basic members and terminals of the relations.

REFERENCES

[1] E.g. V.H. Yngve: Syntax and the Problem of Multiple Meaning, in Machine Translation of Languages, 14 Essays, ed.: W.N. Locke and A.D.Booth, New York, 1955.

Д.Ю. Панов: Автоматический перевод, изд. Ак. Наук, Москва, 1958. 50 p.

[2] V.H. Yngve: Mechanical Translation, Quart, Progr. Report, MIT, 1953.

[3] Gy. Hell: О некоторых характерных чертах алгоритма машинного перевода с русского языка на венгерский, Slavica, Debrecen, 1963.

[4] E.g. K.E. Harper: The position of prepositional phrases in Russian, MT, 1964, aug. 5-10.

I. Panevova: Несогласованное определение точки зрения анализа для машинного перевода, Prague Studies in Mathematical Linguistics, 3. 1966, 219-240 p.

Gy. Sipőczy: The Analysis of Prepositional Constructions, Linguistics, IV. Bp. 1965. 79-92 p.

[5] M. Masterman: Semantic Message Detection for Machine Translation, using an Interlingua, in Proceedings of the 1961 Internat. Conf., Her Majesty's Statt. Office, London, 1962.

K. Spark-Jones: Synonymy and Semantic Classification, Cambridge, Language Research Unit, 1964.

J.J. Katz, J.A. Fodor: The Structure of Linguistic Descriptions, The MIT Press, Cambridge (Mass.), 1964.

F. Kiefer: Some Questions of Semantic Theory, Computational Linguistics, IV. 1965. 71-78p.

[6] P.L. Garvin: A Linguist's View of Language-data Processing, Natural Language and the Computer, ed.: P.L. Garvin, Mecgraw-Hill, 1963. 123.

[7] K. Spark-Jones: Experiments in Semantic Classification, MT, 1965. 97-112 p.

[8] M. Zarechnak: An Applied Radical Semantics, MT, 1965. 90-96p.

[9] It appears that a note should be added to this question. It is beyond doubt that so-called "subject-heading" relying on direct experience constitutes the essential foundations of all dictionaries of definitions. However, it would be too audacious to postulate that also a systematic dictionary of definitions should be based on such "subject-headings" ("object-words"), for these words vary by individual age and language (see B. Russel: An Inquiry into Meaning and Truth, Pelican Book, 66p.). This may perhaps account for that not even the editors of a dictionary of detinitions are consistent in the application of the system of "subject-headings" in the definitions.

[10] See [5], F. Kiefer: Some Questions . . . .

THE ORDER OF THE SYNTACTIC ELEMENTS OF PRINCIPAL SENTENCE IN THE RUMANIAN LANGUAGE DETERMINED BY THE METHOD OF THE THEORY OF GRAPHS

Emese Kis, Elena Comșulea, Ioana Anghel

1. The theory of graphs is an important chapter of modern mathematics and describes the quantitative, relational and formal aspects of different categories of phenomena. It introduces in linguistics a certain precision of the quantitative appreciation, an order and a systematic spirit in the classification of facts. For the Rumanian language S. Marcus and Em. Vasiliu [1] have given such a formal description of the system of consonants.

2. In this exposition the authors propose to separate some quantitative and formal aspects from the complex of phenomena in the most frequent relations among the order of words in the Rumanian language. The material of this examination is to be found, collected and systematized in two earlier papers [2] connected with the order of the syntactic elements in the contemporary Rumanian belles-lettristic prose.

We have studied one thousand principal sentences, extracted from a text [3] composed of twenty thousand words. Whereas the average length of these principal sentences is four (syntactic elements), we had to study all the combinations of the four successive elements, named by us tetragrams. All the syntactic elements occurring in the sentences - except the initial three and the terminal three - took successively the initial place (that is to say: -3), the antepenult place (i.e.: -2), the penult one (i.e.: -1) and the terminal place (i.e.: 0). Any of the places may be a node of these graphs:

$$-3 \longrightarrow -2 \longrightarrow -1 \longrightarrow 0$$

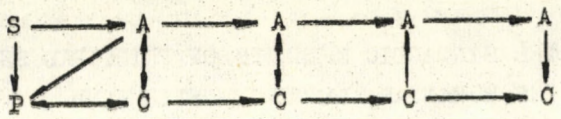
3. Thus the tetragrams were expressed with the help of the theory of graphs, obtaining some new precisions and interpretations concerning the syntax of the Rumanian language. Applying S. Marcus' ideas, by which two problems (one concerning the phonemes, another concerning the grammatemes [4]) are to be resolved by a uniform procedure - we indicate the possibility of transition from a syntactic element towards another one by arches provided with arrows. Consequently any tetragram, realizable from amongst those seven syntactic elements [5], may be represented with a graph similar to the former. The successive elements constitute the extreme points of the arches of the respective graph [6].

In this manner the sentence: "... le pecetluia viața pentru de-a pururea" [7] is in fact the tetragram: CPCC. The successions CP, PC and CC may form the arches of the graph

$$C \longrightarrow P \longrightarrow C \longrightarrow C$$

We tried a systematization according to frequency (see Appendix I) of the 232 tetragrams which realized from the sum of $7^4 = 2401$ possible tetragrams.

A set of all the tetragrams up to the absolute frequency of 27 may be represented as the following centralized cyclic graph Γ:



The cyclomatic [8] number of the graph Γ is 5. This number is to be obtained by the subtraction of the number of nodes (10) out of the number of edges (14) and by addition to it the number of connected components (1), that is: 14 - 10 + 1 = 5. If we start from either of the nodes and we number successively four elements following the direction of the arches, we would obtain any of the tetragrams with a frequency over 27. Accordingly any tetragram with the frequency of over 27 (see Appendix I) may be inscribed as a subgraph of this cyclic graph. If we take a tetragram with the frequency 27, i.e. SCPC, representing a sentence like: "Vale însă dimpotrivă, privi cele spuse" [9], this cannot be inscribed in this centralized graph; even less can be inscribed a tetragram with a frequency 1, like SPCS, i.e. "Divorțul devine și el o problema de clasă." [10].

After the critical frequency of 27 which constitutes anyhow a boundary-mark regarding the frequency of the tetragrams, it is not possible to represent them through a cyclic centralized graph: the cycle will become disconnected. Up to this critical frequency the tetragrams were derived or developed from the SAPC tetragram which represents the typical syntactic order of the Rumain language [11].

It would be interesting to determine the maximal number of the elements which cannot at all be placed together two by two up to the point of this critical frequency. That is to say: if we were to take in any combination two of these elements, they could not stand beside each other up to the point of this frequency. In order to see these combinatorical possibilities we say that two elements, x and y are in relation $x \varrho y$, if the succession xy is possible. That number of elements from which the relation $x \varrho y$ is impossible, is the number of internal stability. Those xy pairs of elements which have no $x \varrho y$ relation among themselves, are named an internally stable set. Up to the frequency of 27 for the elements S and C the relation $x \varrho y$ does not occur: the centralized Γ has an internally stable set; the number of internal stability belonging to this graph is 2. Very near to this freqnency, that is to say up to the frequency of 31 the x y relation does not exist, neither for the A and P elements. Consequently, up to the frequency of 31 the number of the internal stability of the centralized graph is 4, and we have two internally stable sets.

4. Further, any tetragram that has a frequency over 5 (see Appendix I) can be inscribed as a subgraph of another centralized graph which has no internally stable set and its number of internal stability is 1. When we exclude the possibility of crossing the arches, then this centralized graph cannot be represented but in space [12]. The geometrical figure obtained from the centralizing of arches will be a projection in two projective planes of a cube. Such a geometrical figure is convenient to represent the combinations among four variables in mathematical logic [13]. That is to say, up to the frequency of over 5 we find only combinations among C, P, A, S; such syntactic elements that are more frequent than E or V or I. We have no combinations formed

exclusively from E, V, I.

5. We obtain projections in two projective planes of another cube (Fig. 1., graph Δ), if we represent the initial tetragrams which have a higher frequency than 2 (see Appendix II).

In order to inscribe these 42 out of 113 initial tetragrams (see Appendix II), we have introduced a relative regularity in the fixation of the nodes in the graph Δ. On one side of the cube we fix the four syntactical elements of the sentence: S, A, P, C. The predicate has only two nodes, one in the cube, another in the projection, because it cannot occur in a tetragram but once. These two nodes are not adjacent, as the predicate is not to be combined with itself in the frame of the same sentence. On the two edges of the cube we put SA, SC and CA. On one side of the projection a complement corresponds to three attributes and an attribute to three complements. On the two antipode-edges we have AA and CC. The succession CCCC is not forming a cycle, so that every node C may have the possible greatest free degree and grade for the other elements. In fact in this graph the cycles very seldom realize as circuits.

Whereas there are reflexive and symmetric arches, neither graph Γ, nor graph Δ can be regarded chromatics relating to a given coloration.

It would be interesting to compare the Γ and Δ graphs. In the graph Δ in distinction from Γ the number of realized edges is 31, the number of nodes is 16 and there is but 1 connected component; therefore the cyclometrical number of the graph Δ is $31 - 16 + 1 = 16$.

6. Similarly it is possible to compare these two graphs from the point of view of the relation $x \varrho y$. Declining from Γ in the graph Δ the relation $x \varrho y$ occurs among all the elements S, A, P, C. Consequently we have no internally stable set.

If we compare this centralized graph Δ with a graph representing all the initial tetragrams (see Appendix II and Figures 2-30), we may establish that the elements E, V, I form an externally stable set in front of the elements of the graph Δ. That is to say: any element, not belonging to graph Δ, but belonging to the totality of tetragrams, can form the extreme node of an arch which starts from an element of the graph Δ (see Figures 6, 7, 15, 16, 21,22, 23, 28).

Viewed synoptically, the arches of this graph Δ are segments of a spiral which is "twisting" round the predicate, touching it or departing from it at smaller or greater distances.

It is to be observed about graph Δ (see Figure 1) that of the first elements of the sentence, C and the diagram CC have the highest frequency. This fact is explainable by the circumstance that in the Rumanian language the subject is the most redundant element regarding the grammatical information (it may be expressed by the verbal complex or implied in the sentence or it may be simply left indeterminate).

The sentences are very often composed only of a predicate, or of a predicate and its complements, or of a predicate, its complements and the attributes of the complements. The fact that the graph Δ contains 16 nodes must be interpreted in the following way: in the principal sentence any of the elements (except E, V, I) can stand before the predicate or before another sentence element which precedes the predicate.

7. Comparing this statement with assertion of S. Marcus and Em. Vasiliu [14] concerning the adherence of the consonants to the vowels, we observe an isomorphism betseen the structure of initial consonants and their adherence to the vowel on the one hand, and the structure of the elements of the sentence and their adherence to the predicate in the initial tetragrams on the other.

The isomorphism among the structure of the syllable and the structure of the sentence has been established by J. Kurylowicz [15] and was mathematically formulated by S. Marcus [16]. To this isomorphism concerning a relation of domination valid in all languages, we may add another isomorphism concerning an adherence relation to the dominant element which is for the present studied in the Rumain language.

We have two elements of the sentence: x and y. We would say: x adheres to the predicate from left rather than y, in the case when the sequence of yxP may be realized in the Rumanian language. We shall write: $x \xrightarrow{P} y$. For example C adheres better from left to P than S, because the digram SC is more frequent than CS. We shall write: $C \xrightarrow{P} S$. The element of sentence with the greatest adherence from left to the predicate is C, whereas with the least frequency is E which has 0 adherence. In the centralized graph Δ any subgraph containing a way with three arches can be noted in the following manner: the last node is equal with 0, the penult with -1, the antepenult with -2, and the first node is equal with -3.

In the study of the adherence of sentence-elements to the predicate we only have to regard those subgraphs that have the predicate with the note 0 (see Figures 4, 10, 19, 25). These graphs, taking into consideration the frequency of the adherence of the syntactic elements to the predicate, appear as follows: in node -1 (which is nearest the predicate): $C \xrightarrow{P} S$, $S \xrightarrow{P} A$; in node -2: $C \xrightarrow{P} S$, $S \xrightarrow{P} A$; in node -3: $A \xrightarrow{P} C$, $C \xrightarrow{P} S$.

As $C \xrightarrow{P} S$, $S \xrightarrow{P} A$, we observe that the mediation of C gives a better adherence to the predicate than S or A. In this sense, concerning the adherence to the predicate, S is dominated by C, consequently the most frequent digrams are CP (see Appendix I). In the sentences with two elements the most frequent type is the type CP (absolute frequency 62); in the sentences with three elements those sentences have the highest absolute frequency (69) which contain the combination CP.

8. We may draw the conclusions:

The isomorphism between the structure of adherence to the vowel in the frame of the syllable and the structure of the adherence to the predicate is relative. The difference is explicable also by means of the reduced number of the functional unity dominated by the predicate in the sentence (3, i.e. 4) and the increased number of the consonants (20).

This fact is illustrated by the reduced number of subgraph-circuits, that is to say, by the graphs which are on the same level of the geometrical figure. Consequently the digrams - combinations among two elements - have a higher importance than trigrams or tetragrams. Regarding the regular order presented in the grammars: SAPC, we may generally establish that SA is more frequent than AS, PC is more frequent than CP. It follows from this that the fixation of the correspondence among the combinations of two per two elements (clos-

ed digrams and open ones) is urgently necessary for the construction of an algorithm for machine translation.

From those stated above there follows a typological conclusion: the syntactic order of the Rumanian language formulated in this manner is comparable with the order of sentences of the Russian language formulated mathematically [17]. In contrast with the Russian language having a linear syntactic order, the Rumanian language has a complex syntactic order which after a certain frequency of the combinations among the syntactic elements cannot be represented but in space.

REFERENCES

[1] S. Marcus, Em. Vasiliu: Matematica şi fonologie. Teoria grafelor si consonantismul limbii romîne, in "Fonetica şi dialectologie", III. (1961), p. 15-55.

[2] Joana Anghel, Elena Comsulea, Emese Kis, Ioan I. Stan: Observaţii relaţionale şi statistice privind topica propoziţiei principale, in SCL, XI. (1966) (sub tipar); idem, Cu privire la topica complementului în propoziţia principala, in CL, X. (1965), nr. 1, p. 209-220.

[3] Tudor Arghezi: Cu bastonul prin Bucureşti (Bucureşti), Editura pentru literatura, 1961, p. 112-122, 135, 148, 157;

Eugen Barbu: Soseaua Nordului (Bucureşti), ESPLA, 1959, p. 381, 526;

Mihai Beniuc: Pe muche de cuţit (Bucureşti), ESPLA, 1959, p. 25, 44, 263;

Geo. Bogza: Meridiane sovietice (Bucureşti), ESPLA, 1953, p. 9, 11, 50,133;

Eusebiu Camilar: Negura (Bucureşti), Editura de stat, vol. I, 1949, p. 20, 131, vol. II, 1950, p. 193;

G. Calinescu: Scrinul negru (Bucureşti), ESPLA, (1960), p. 285, 541, 599 ;

V. Em. Galan: Baragan (Bucureşti), Editura tineretului, 1959, vol. I, p. 313, vol. II, p. 129;

Francisc Munteanu: Statuile nu rîd niciodata (Bucureşti), Editura tineretului, 1957, p. 89, 343;

Ion Pas: Lanţuri (Bucureşti), ESPLA, vol. I, (1950), p. 13;

Camil Petrescu: Un om între oameni (Bucureşti), Editura tineretului, vol. I, 1953, p. 399, vol. II, 1955. p. 482;

D.R. Popescu: Umbrela de soare (Oradea), Editura tineretului, (1962), p. 214, 309;

Titus Popovici: Setea (Bucureşti), ESPLA, (1958), p. 179, 319;

Marin Preda: Risipitorii (Bucureşti), Editura pentru literatura, (1962) , p. 127, 391;

Mihai Ralea: In extremul occident, Note de drum din Antile, California , Canada (Bucureşti), ESPLA, (1955), p. 26, 76, 110;

Vasile Rebreanu: Casa (Bucureşti), Editura pentru literatura, 1962, p. 85, 95;

M. Sadoveanu: Mitrea Cocor, Bucureşti, ESPLA, 1949, p. 39, 131;

Zaharia Stancu: Radacinile sînt amare (Bucureşti), ESPLA, (1958), vol. I, p. 157;

Radu Tudoran: Dunarea revarsata (Bucureşti), Editura pentru literatura, 1961, p. 190, 419;

Ion Vlasiu: Drum spre oameni (Bucureşti), Editura pentru literatura, 1961, p. 135, 319;

Tiberiu Vornic: Sub pajura împaraţiei (Bucureşti), ESPLA, (1954), p. 191, 314;

[4] Cf. Solomon Marcus: Lingvistica matematica. Modele matematice în lingvistica. Editura didactica şi pedagogica, Bucureşti, 1963, p. 50-51;
In morphology this method was applicated by Ján Horecky: Moftemetická struktura slovenciny, Vydavatelstvo Slovenskej Akademie vied, Bratislava, 1964.

[5] S = subiect, P = predicat, A = atribut, C = complement, E = element predicativ suplimentar, V = vocativ, I = interjecţie.

[6] S. Marcus, Em. Vasiliu: op. cit., p. 15 ş.u.

[7] Eusebiu Camilar: op. cit., p. 20.

[8] S. Marcus, Em. Vasiliu: op. cit., p. 18.

[9] Marin Preda: op. cit., p. 127.

[10] Mihai Ralea: op. cit., p. 110.

[11] "Ordinea cea mai obişnuita în limba româna într-o propoziţiel principala dezvoltata, cu diferite parţi de propoziţie, este urmatoarea: subiect - - atribut - predicat - complement direct şi indirect - complemente circumstanţiale." Gramatica limbii române", vol. al II-lea, Ediţia a II-a revazuta şi adaugita, Editura Academiei Republicii Populare Române, Bucureşti, 1963, p. 428.

[12] I. Anghel, E. Comşulea, E. Kis: Topica propoziţiei principale prin prisma teoriei grafelor, Colocviul de teoria funcţiilor convexe cu aplicare la calculul numeric, Cluj, 1 - 5 iulie 1965.

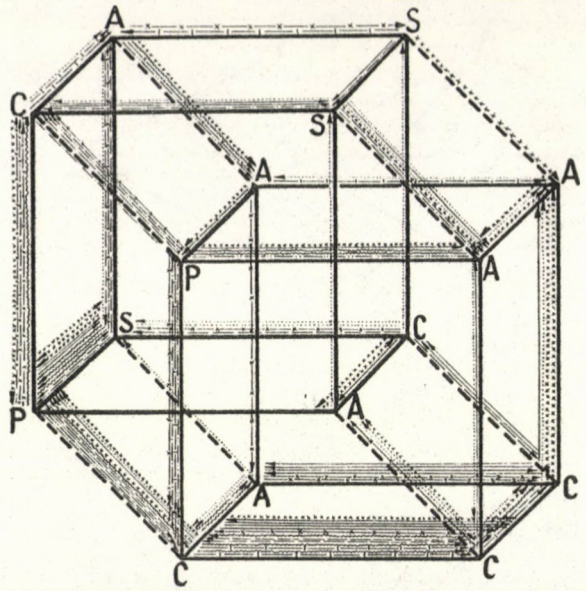
[13] Varga Tamás: Matematikai logika, Tankönyvkiadó, Budapest, 1962.

[14] S. Marcus, Em. Vasiliu: op. cit.

[15] J. Kurylowicz: La notion de l'idomotphidmr, in "Travaux du Cercle Lingvistique de Copenhague", V. (1949), p. 8.

[16] S. Marcus: Aspecte ale modelarii matematice, in "Studii şi cercetarii lingvistice", XL. (1963), nr. 4, p. 498, ş.u.

[17] Cf. P. Adamecz: Slovosled v rustine, Praha, 1963;
E. Pauliny: Slovosled a aktuálné vetné clenenie, in "Slovenska rec", XVI. (1950-51), p. 171-179.



S = subiect. P = predicat. C = complement. A = atribut.
Arc inițial frecv. 3-4 ———, 5-11 — — —, 12-30 —·—·—
Arc medial frecv. 3-4 ·····, 5-11 —·—·—, 12-30 —·—·—
Arc terminal frecv. 3-4 ·····, 5-11 —×—×—, 12-30 —|—|—

Fig. 1.

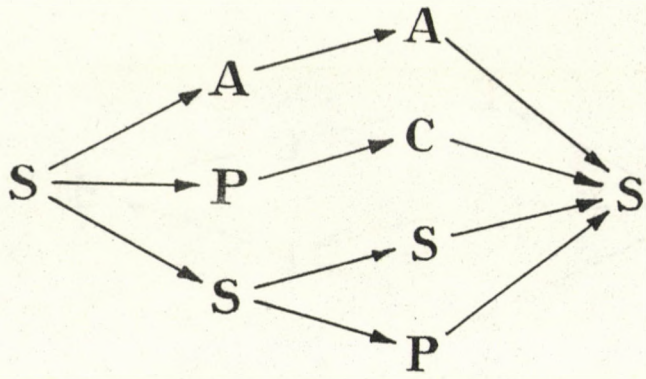


Fig. 2.

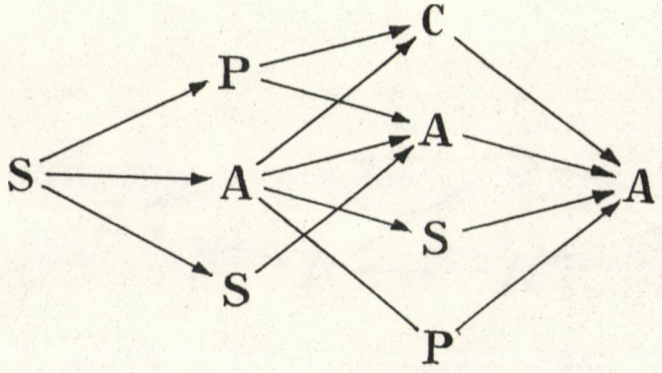


Fig. 3.

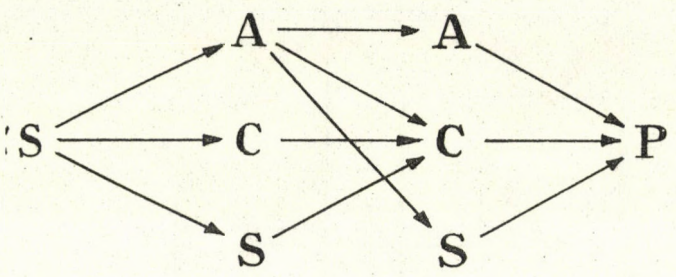


Fig. 4:

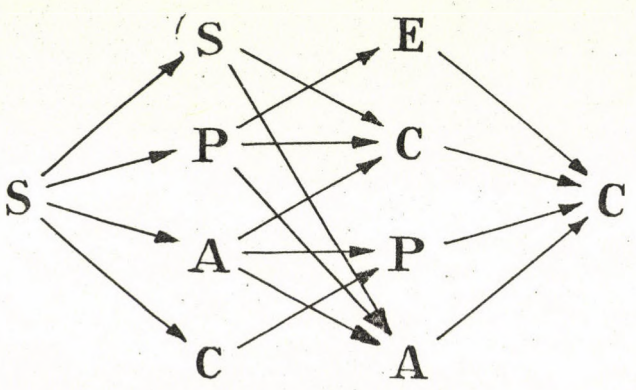


Fig. 5.

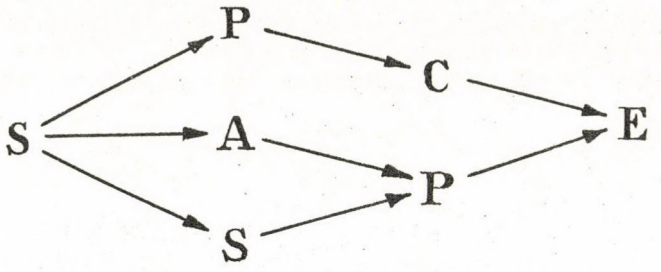


Fig. 6.

Fig. 7.

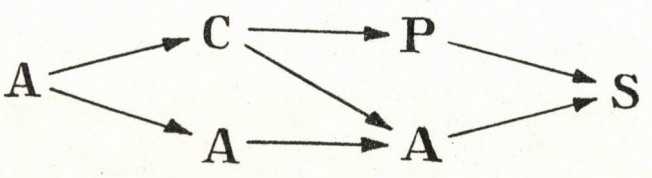


Fig. 8.

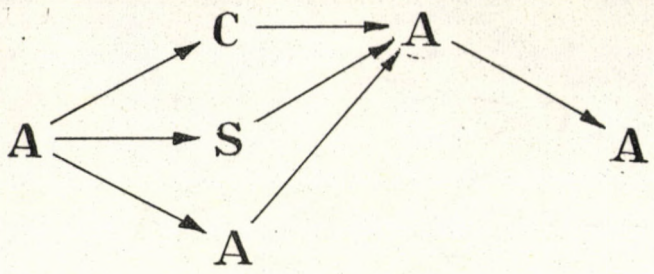


Fig. 9.

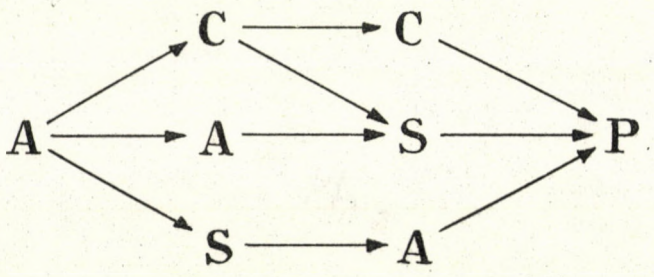


Fig. 10.

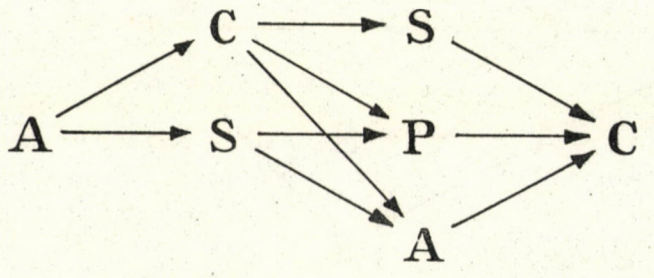


Fig. 11.

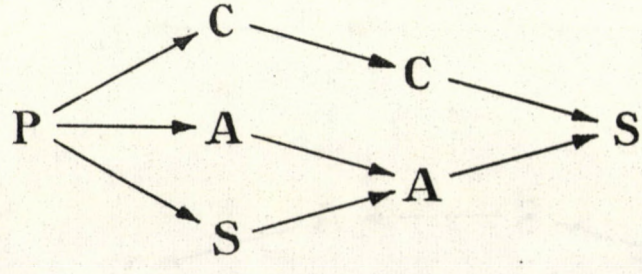


Fig. 12.

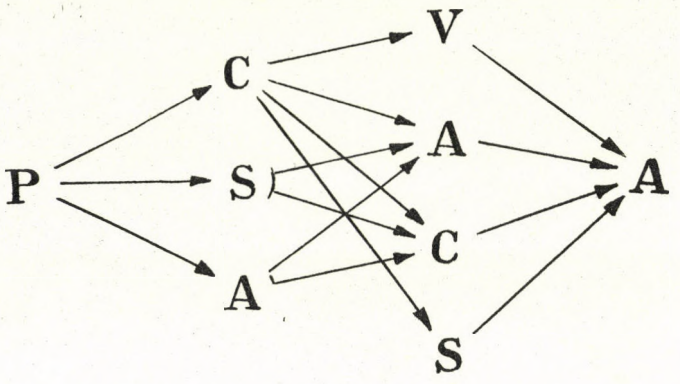


Fig. 13.

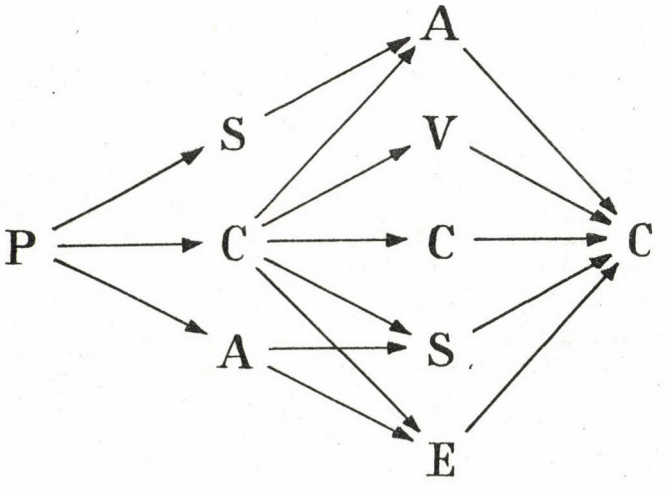


Fig. 14.

Fig. 15.

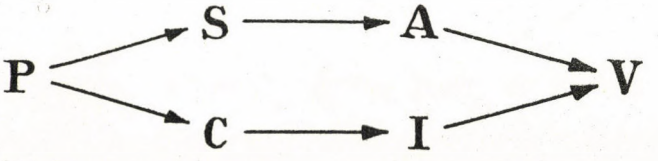


Fig. 16.

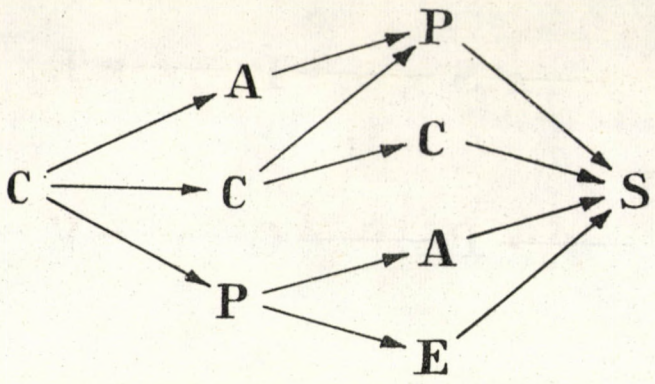


Fig. 17.

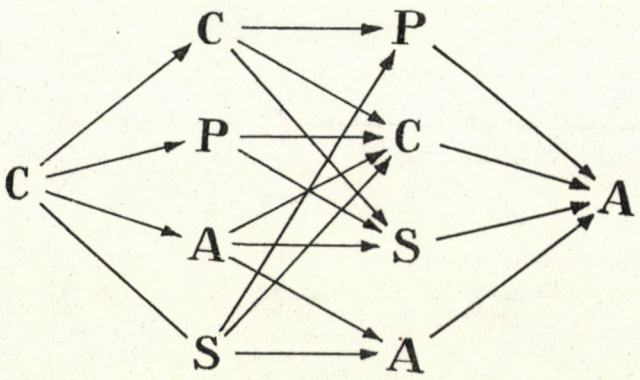


Fig. 18.

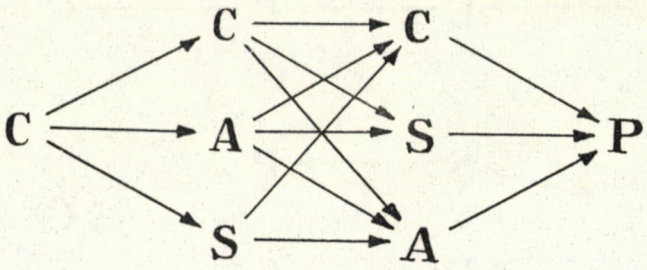


Fig. 19.

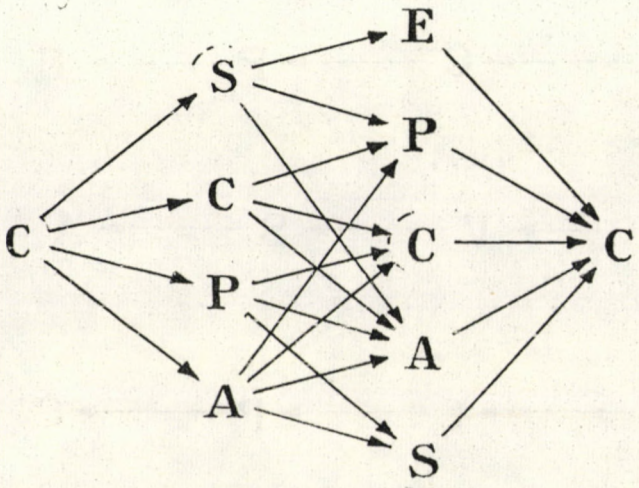


Fig. 20.

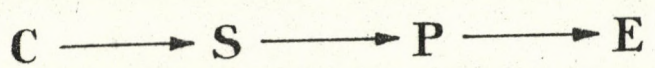


Fig. 21.

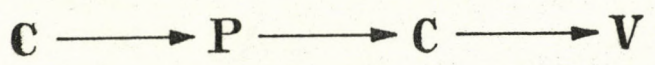


Fig. 22.

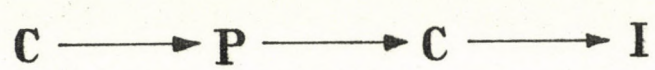


Fig. 23.

Fig. 24.

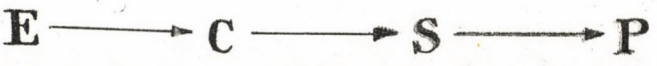


Fig. 25.

Fig. 26.

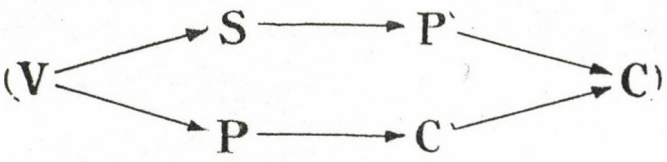


Fig. 27.

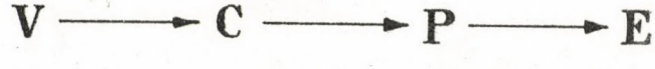


Fig. 28.

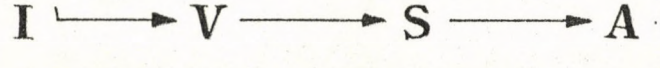


Fig. 29.

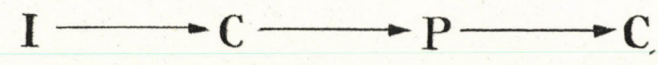


Fig. 30.

APPENDIX I
(All the tetragrams)

132	94	62	60	56	47	45	43	42	40	39
P---	AAAA	CP--	CAAA	PC--	PCCA AACA	ACAA	AAAC	PCCC	SAAA	SP--
38	37	36	35	33	32	31	28	27	25	24
CAAC	PCAA	CCAA	CPCA	PCAC	CACA	SAPC CCCC	SPCA APCC	SCPC CPC-	CCCA	SPCC
23	22	20	19	18	17	16	15	14	13	12
AAPC	APCA	ACAC PS-- SPC-	ACPC	CCPC	CCAC ASPC PSAA ACCA	CAPC AACC	AAAP CPSA	SAAP CASA ASAA	SCP-	CACC CCP-
11	10	9	8	7	6	5	4	3	2	2
SACP CSPC AACP	CPS- SAP-	ACCC CCCP CPSC SASA	C--- CSP- ASAP CACP PSCA	ASAS PACA CSAA PSC- APSA	AAAS SAAC ACAP CASP	CCAP CAAP CCAS CAEC AASA PCA- PV-- AASP CCSA CSAP CCSP CSCP ACPS CAPS CPAC	CCCS CECC SS-- AAPS CCPS CCEC ACCP SAAS PCCE PCAS PCSA SPAA AECA PAAA PSAC SACA SCCP S---	SPAC CAAS SSCP ECCC IP-- PCCS CSPE PACC PAC- ECAC APCS SAPA ACAS PAA- CCPA CSAC PSCC PAAC CC-- PCV- SACC SCAC	CSCA VPCC CPEE AECC PCEC PCIV PAEC PSA- PE-- SPCE SSSC ECCE ECCA ECAE ECEC EECC CASC SASP SSAA VPC- SCPA SPEC CSPA CECA	CPV- PCAE PECA ECAA SCAA SA-- SPE- VP-- ICP- ACPE AAPA PASC PSAS PSSA CPCV CPAS ACSP ASCA APAA PCS- SCAP SPA- SSSA VV-- PCVA

1	1	1	1	1	1	1	1	1	1	1
AASS	PSCE	CCIV	CCSS	CEAA	APCE	PCEE	PI--	ECSP	SSCC	SSAS
ASSP	PVC-	CPCI	CCEA	CEAC	APEE	PCEA	SACS	EACC	VSAP	SSSS
CACV	ACAE	E---	CAIV	CEEA	AP--	PACE	SASS	EAP-	VSA-	SCCA
CPES	APEC	CPEC	CPCS	ACCS	ASCP	PAAS	SASE	EPCA	CCA-	SAPV
CVCA	PSV-	APAC	CPAE	ACPA	ASP-	PAS-	SPCV	VCPE	CAP-	PEA-
AC--	SAPE	APSS	CPVA	ACSS	ASSS	PSAV	SSPC	VSP-	ACP-	ISP-
PCVC	SPSC	SAAE	CSSS	ACEC	ASEC	PEES	SEPC	VSPC	CCAE	CS--
PCE-	SPV-	SSP-	CSSP	AACS	AS--	PEEC	SECE	IV--	CSCV	CV--
PEC-	SSPS	CCC-	CSEC	AAPE	AECE	PEEE	ECCP	ASCC	ACSC	V---
ASA-	CCE-	CCSC	CECE	AASC	PCAI	PVAA	ECPC	SPCS	APVA	I---

APPENDIX II

(Initial tetragrams)

30	23	21	18	16	14	12	11	10	9	8
CPCC	SAPC	CSPC	SPCA	PCCC	SPCC	PCAA	SAAP	SACP	CCPC	CCCP
			SAAA			PCAC	CPSA	PCCA	CAAC	
			CPCA							

7	6	5	4	3	2	1	1
CSPC	CAPC	CCCC	CACA	CCSA	CSCA	PCCS	PSAS
ASPC	CPSC		CCAC	CCSP	SPAA	PCSC	SAPA
			CASA	CSCP	CPAC	SPCS	SPCV
			CSAA	SAAC	VPCC	SSCC	SSCP
			CCCA	CCPS	CCCS	SASP	SSAA
			PSAA	ACAA	CPAC	IVSA	SSAC
			PSCA	PACA	ACAS	AASP	ECSP
			CASP	CACP	PCIV	ACSP	EPCA
				CAAP	SPCE	SSSS	EECC
				CACC	CASC	ACSC	VCPE
				CSAP	ASAA	AAAA	VSPC
				SCCP	PSAC	CAPS	CPCI
					PAAA	CAAA	CPES
					PASC	CPAS	CSPE
					SACA	CSPA	ACPS
					SAAS	CSEC	PCCE
					SASA	ACPC	PCSA
					CSAC	ACAC	PCVA
					CCPA	AAAS	PCVC
					CPCV	ASAP	SAPE
					ACCP	PCEC	SPEC
					ASAC	PAAS	SSPS
					CCAP	PAEC	SSPE
						PSAV	ICPC

ОДИН ВОЗМОЖНЫЙ СПОСОБ СНЯТИЯ ОМОНИМИИ ОСНОВ ПРИ АВТОМАТИЧЕСКОМ АНАЛИЗА РУССКОГО ТЕКСТА В ЦЕЛЯХ МАШИННОГО ПЕРЕВОДА

А. Людсканов, Е. Паскалева

Омонимия средств на любом уровне языковой структуры всегда представляла собой одну из основных трудностей при автоматизации любых операций над текстами, материализованными в форме естественных языков (L_1^N) в том числе и при машинном переводе /МП/. Методы автоматического снятия омонимии зависят, с одной стороны, от уровня рассматриваемых единиц /основы, т.е. лексемные единицы; морфемы, т.е. морфологические единицы; конфигурации, т.е. синтаксические единицы/ и, следовательно, от типа словарных единиц, организации словаря и количества и вида словарной информации /I/, а, с другой - от принятого типа анализа и общей концепции его организации.

Настоящее сообщение ставит себе целью описать способ снятия омонимии [1] лексемных единиц, принятый в разрабатываемой группой Машинного перевода и Математической лингвистики при МИ БАН алгоритме МП русских математических текстов на болгарский язык.

Общая концепция автоматического анализа, принятая нашей группой, и от которой существенно зависит тип и способ реализации разрабатываемого алгоритма и, следовательно, способ снятия омонимии лексемных единиц, сводится к следующему. В настоящее время господствующим направлением в области автоматического перевода является разработка т.н. "смыслового" МП /срв. напр. Л.1, 2 и 3/. При этом, выделяя в L_1^N отдельные, связанные отношением форма - содержание, уровни /l_1 - графемический, l_2 - семический, l_3 - синтаксический, l_4 - семантический/, цель анализа мы усматриваем в последовательном приведении всех средств отдельных уровней на уровень l_4, т.е. "извлечение смысла" и представления его в данной единой форма записи [2], в качестве основы синтеза.

Исходя из предложенного в /Л.5/ понятия "необходимая переводная информация" - I/T/ и из учета структурной близости славянских языков, нам представляется более рациональной следующая организация анализа. Вместо того, чтобы сразу трансформировать все элементы всех уровней на уровень l_4, извлекать всю возможную I/T/ на данном уровне и данным типом анализа, и только в тех случаях, когда это невозможно, переходить к следующему уровню.

На базе такой концепции анализа в качестве основы нашего алгоритма принят поморфемный лексико-морфологический анализ с учетом словесного окружения, и предполагается введение элементов операционного синтаксиса для неразрешимых на этих уровнях проблем /напр. проблема т.н. "членной формы", т.е. артикля и др./. В связи с этим, используя вариант предложенной И.А.Мельчуком процедуры сегментации /срв. Л.6/ и исходя из corps de texte в 62 000 словоформ, было выделено 70 словоизменительных и 26 словообразовательных единиц, т.е. суффиксов /включенных в морфологический словарь - D_2/ и 1682 основных морфем /включенных в лексемный словарь - D_1/. Поскольку уменьшение объема словарной единицы ведет к увеличению омонимии, в D_1 199 омонимичных основ.

Исходя из принятой концепции разрешать все возможные случаи на данном уровне, не переходя к более глубокой структуре и более сложному типу анализа, мы попытались построить схему снятия этой омонимии на морфологической базе, т.е. анализа остатка (d_o) не прибегая к контексту и синтаксическому анализу. При неомонимичных суффиксах однозначное определение принадлежности данной основы при данной текстовой единице к данному классу не представляет трудности /напр., при анализе словоформы вероятн-ый на основе полученных дан-

ных о принадлежности суффикса ый к классу прилагательных, мы сможем сразу снять омонимию основы вероятн-/. Известные затруднения возникают, однако, при анализе ν , образованных от омонимичных S и нескольких суффиксов, каждый из которых может фигурировать в разных словесных классах /напр. суффикс -ост- в вероятн-ост-ь и вероятн-ост-н-ый/. В этих случаях машине пришлось бы проверять все словообразовательные суффиксы вплоть до окончания, которое со своей стороны в большинстве случаев тоже является омонимичным и не дает необходимой I. На первый взгляд эта одновременная омонимичность и основы, и суффиксов ν ведет к заколдованному кругу, единственный выход из которого ведет к анализу на синтаксическом и семантическом уровне. Однако в процессе практической работы мы пришли к следующей констатации, которая позволяет построить определенный вариант снятия омонимии на этом уровне.

Суффикс, находящийся непосредственно перед окончанием /наличным или нулевым/, несет однозначную информацию о принадлежности данной основы к определенному словесному классу. Тот же суффикс, находясь в другой позиции в составе ν , может принадлежать к другому классу /напр. суффикс -н- в словоформах квадрат-н-ый и квадрат-н-ост-ь/. Приведенная ниже таблица наглядно иллюстрирует зависимость между позицией данного элемента в морфематической структуре ν и принадлежностью данной основы к определенному словесному классу. Кроме этой I, предпоследний суффикс может нести и однозначную I о некоторых других грамматических данных /напр., род при S , вид при V и . пр./.

ТАВЛИЦА [3]

суф.	Информация о принадлежности к данному словесному классу, которую несет суффикс, находясь на предпоследнем месте.	Другие словесные классы, в которых данный суффикс может фигурировать, не находясь на предпоследнем месте.	Другие грамматические характеристики, необходимые для снятия омонимии.
а	V /организ-ов-а-л/	S /организ-ов--а-нн-ост-ь/ S /квадрат-а / A /хорош-а/	
ва	V /созда-ва-ть/	S /созда-ва-ни-е / A /созда-ва-ем-ый /	несов. в.
ев	A /нул-ев-ой/	S /пальц-ев/	
ем	A /наблюда-ем-ый/	S /дел-ени-ем/ V /наблюда-ем /	
енн	A /существ-енн-ый/	S /числ-енн-ост-ь/	
е	V /завис-е-л/	S /числ-е/ A /лучш-е/	
им	A /дел-им-ый/	V /привод-им/	
и	V /подчин-и-ть/	S /математик-и/ A /хорош-и/	
нн	A /организ-ов-а-нн--ый	S /орагниз-ова-нн-ост-ь/	

ов	A /числ-ов-ой/	S /квадрат-ов/ V /организ-ов-а-ть/	
ост	S /вероятн-ост-ь/	A /вероятн-ост-н-ый/	ж.р.
тел	S /дел-и-тел-ь/	A /вычисл-и-тел-ь-н-ый/	м.р.
ут	A /замкн-ут-ый/	S /замкн-ут-ост-ь/ V /замкн-ут/	
у	V /толкн-у-ть/	S /числ-у/ A /эт-у/	
ыва	V /доказ-ыва-ть/	S /доказ-ыва-ни-е/	несов. в.
я	V /раздел-я-ет/	S /земл-я/	

При использовании I, которую несет предпоследний суффикс для снятия омонимии основ, особую трудность представляет сигнализация нулевого окончания. Ошибочная сигнализация ведет к неправильному определению морфематической структуры слова и оттуда – к ошибочному определению принадлежности к данному словесному классу /напр., в словоформе организ-ов-а-н- #, если нулевое окончание /#/ не сигнализировано, предпоследним суффиксом выступит суффикс -а-, и словоформа будет отнесена к V вместо к A/. Несмотря на то, что мы не в состоянии дать здесь подробное описания, мы должны отметить, что правильная сигнализация нулевого окончания обеспечивается следующим способом: в D_1 вводится специальный индекс, указывающий на способность данной основы присоединять нулевой суффикс в некоторых из своих форм. В D_2 вводится индекс о способности данного суффикса присоединять нулевой суффикс. Путем последовательного накладывания и сопоставления этих ограничительных условий получается необходимое фильтрование [4].

Л И Т Е Р А Т У Р А

1. И.А. Мельчук, А.К. Жолковский: О возможном методе и инструменте семантического синтеза, Научно-техническая информация, № 6, 1965 г.
2. Я. Паневова, П. Сгал: Структурна и математическа лингвистика в ЧССР, Език и литература, № 6, 1964 г.
3. А. Людсканов: За "лингвистическия модернизъм", "дехуманизацията" на езикознанието и точните методи на изследване, Списание на Българската Академия на науките /под печат/
4. Сб. Машинный перевод и прикладная лингвистика, вып. 8, М. 1964 г.
5. А. Людсканов: Основи на теорията на машинния превод с оглед на руско-българския МП, Годишник на ФФ на СУ, С., 1964 г.
6. И.А. Мельчук: Морфологический анализ при машинном переводе, Проблемы кибернетики, вып. 6, М., 1961 г.

[1] Под омонимией мы будем понимать здесь т.н. грамматическую омонимию в лексемном словаре, т.е. случай, при котором от одной словарной единицы /основы - s/ могут образовываться словоформы /v/, принадлежащие к разным словесным классам - S /существительное/, A /прилагательное/, V глагол и т.д. В лексемном словаре такие основы снабжены несколькими информационными рядами, и цель схемы "Омонимия" сводится к однозначному выбору того информационного ряда, который соответствует принадлежности данной ν, образованной от данной S к данному словесному классу.

[2] Например, в форме т.н. "семантических множителей" /срв. напр. Л.4/ или своеобразного языка "Basic" /срв. напр. Л.1/ или отдельного языка-посредника со своей порождающей грамматикой /срв. напр. Л.2/ и пр.

[3] В таблица приводятся только омонимичные суффиксы.

[4] Известные затруднения создает и аглютинативный характер возвратной частицы -ся /сь/, которая, присоединяясь к ν, изменяет порядок морфологических элементов. Поэтому, при наличии возвратной частицы, анализ структуры ν проводится без учета этой частицы, причем возвратное значение словоформы восстанавливается при синтезе.

CONCERNING THE FORMAL DESCRIPTION OF LANGUAGES FOR TRANSLATION WITH ELECTRONIC COMPUTER [1]

J. Máthé, P. Schweiger

0. The nature of the relationship between language and speech makes the potential sign (the language sign) actualize only in the sign of speech. Hence the impossibility of a direct study of the language phenomenon; the indirect way, mediated by research, leads in one way or another to the idea of modelling. Deliberately or not, linguists have only modelled the speech facts from the point of view of those of the language.

Applied linguistics, mechanical translation requiring a greater formalization and a more accurate systematization, have put the problem of a consistent use of modelling technique with the whole mathematical and logico-mathematical apparatus; the logical models, some with unrealized nodes, topological models, statistical models, models based on the set-theory and so on .

On the basis of our preliminary investigations we have come to the conclusion that the most adequate model for the realization of mechanical translation by means of an intermediary language is the logical model. This working assumption begins to show its accuracy more and more on the basis of the concrete investigations of the language facts.

0.1. Starting the research of the material of the three languages: Rumanian, Russian and Hungarian, we have succeded to work out some methods and principles which later on proved to be accurate. We divided the multitude of words of a language into two fundamental submultitudes: the submultitude of uninflected words (prepositions, conjunctions etc.) and the submultitude of the inflected words. The latter have been treated on the basis of their capacity of paradigmatic inflexion. In the process of inflexion we distinguish two elements: the root and the affixe-morpheme, which fulfil the following conditions:

a) the root remains unchanged during the inflexion (except the roots which contain letters with variable value, for example: Rum.: brad-brazi; Rus.: ухо-уши ; Hung.: tó-tavat etc.),

b) each root presents a probabilistic active valency with a great number of affixe-morphemes,

c) the morpheme is different from one circumstance to another and is distributively and probabilistically determined,

d) the affixe-morpheme may have a passive and probabilistic valency compared with diverse roots,

e) for particular cases the affixe-morpheme can be equal with a void multitude.

The above criteria proved to be sufficient for making up the flexional types in the limits of the word classes. The types are made up on the basis of paradigmatic identity of the words of a certain class. We have considered that these classifications may be founded on some intuitive and empiric definition which are verified in the practice of elaborating the algorithm. So the "word" was defined as the series of letters uninterrupted between two blanks and determined by its appartenance to a certain type of flexion or the submultitude of uninflected words.

0.2. The system modelled on the basis of the above principles is probabilistically determined by the relation potential-actual. So we are thinking of the defective nouns for number, which are placed in the flexional system of the nouns of the same type with a complete paradigm,or the Rumanian adjectives the relation of which

$$R \quad M_{ms} \qquad\qquad R \quad M_{fs}$$

$$R \quad M_{mp} \qquad\qquad R \quad M_{fp}$$

is sometimes achieved by

$$R \quad M_{mp} \qquad = \qquad R \quad M_{fp},$$

or the absence of one of these forms. Here we must also mention the case of the Hungarian nouns which do not make actual their potential valency for the 714 inflected forms.

0.3. A fundamental requirement for the mechanical translation according to "grammar" is the consistent solution of the so called exceptions (in declension, in conjugation etc.). As a result of our working principles the "exceptions" have found their place in this system. For mechanical translation it is indifferent wether a certain type of inflection contains one or several thousand words. But the types that contain a single word definitively solve the problems of the "exceptions" by placing them into a system.

0.4. A new notion of the grammar that underlies the translating algorithms is considered to be the letters with variable values. Instead of etymological explanations that do not allow the formalization of their treatment, the letters with variable values permit a good transcription by means of logical orders. For example:

verd - e $\sim$ verz - i

$d_x = (d,z)$, or in Russian

ден' - ф $\sim$ дн' - а

$e_x = (e, \emptyset)$, or in Hungarian

irodalom - $\emptyset$ $\sim$ irodalm - at

$o_x = (o,\emptyset)$

1.0. The elaboration of the algorithm for mechanical translation begins by writing some mechanical dictionaries consisting of lists of word classes and inflexion types for a certain language; the further parts are:

a) the blocks of independent analysis,

b) the translating orders,

c) the blocks of synthesis.

1.1. In our conception the notions "root" and "affixe-morpheme" seldom coincide with the respective notions of the traditional grammar books.

$$w = R + M = (R' + Suf.) + Term.$$

1.2. As we have shown above, in the list of words we find the roots of certain languages arranged in a certain order as well as certain necessary informations; in the cell of the external memory, a root will be written in the following way: the address, the root, the grammatical information, the address of the corresponding word (words) in the target language.

Parallel to the lists of roots we use lists of affixe-morphemes (endings).

These lists contain all the affixe-morphemes of the language. Here is what will be written in the memory cell: the address, the morpheme, the information concerning the class of words and the passive valency of the given affixe-morpheme.

The process of independent analysis consists exactly of the correlation between the information from the list of roots and that of the list of affixe-morphemes and the removal on the basis of the binary microtextual analysis of the redundant information.

2.0. Considering the Rumanian language we worked out the basis of the binary microtextual system and described a list of about 35 types of word groups made up of two or three units. In Rumanian on the basis of "The Dictionary of Contemporary Literary Rumanian" [2] the first 6000 nouns divided into 80 types of declension were worked up. The material so processed is characterized by 21 affixe-morphemes. Each word of the list comprises grammatical information concerning whether it presents root homonymy or not, whether in the plural it has a meaning other than in the singular, and its government.

2.1. The list of Rumanian adjectives of 7.892 items distributed into 60 flexional types were set up on the basis of the same dictionary. The Rumanian adjective is characterized by 63 initial roots in the nominative case.

2.2. In Rumanian there exist 14 basic (simple) cardinal numerals, 17 compound numerals and 72 complex numerals.As examples of simple numerals we give: trei (three), patru (four), zece (ten) etc., compound numerals: unsprezece (eleven), douăzeci (twenty), treizeci (thirty) etc.,complex numerals: douăzeci şi unu (twenty one), treizeci şi doi (thirty two) etc.

3.0. The Russian material was worked up relying on the same principles as for the Rumanian one. The lists of words were set up using "The Dictionary of the Russian language" of S.I. Ožegov [3].

3.1. In Russian there exist 14.000 nouns distributed into 76 flexional types. These nouns are characterized by 40 affixe-morphemes. We have also to mention that there were set up exact tables of the letters with variable values in the declension.

3.2. Counting the Russian verbs from the above mentioned dictionary we have established 80 types containing 9690 words. In the conjugation of the verbs 42 groups of letters may appear with variable values.The list of affixe-morphemes comprises 69 forms. The list of the Russian nouns as well as that of the Russian verbs provides grammatical information of the same type as in the case of the corresponding Rumanian word classes.

3.3. By elaborating the lists that comprise the adverbs, pronouns and particles, the following results were obtained: 1271 adverbs existing in the above-mentioned dictionary are distributed into 30 types; there exist 52 pronouns classified according to their grammatical information. We also classified 58 particles according to their grammatical information.

3.4. The list of the Russian prepositions comprises 61 words together with a large body of information about homonymy (the ability of the given preposition to be used with different cases), government and left and right operators. These operators contain the verifying orders of the word or words preceding the propositions or standing after the preposition. By means of these

operators correlated with the grammatical information of the word having an active valency related to the given preposition, as well as the grammatical information of the noun which has a passive valency, the actual function of the preposition is determined.

4.0. In establishing the types of declension and conjugation of the Hungarian nouns and verbs we set out from the same principles. As in the case of the Rumanian and Russian language, the root is considered that part of the word which remains unchanged during the process of inflexion and the affixe-morpheme the part of the word which is modified. This covers not only the simple words but also the derived and compound words. For example: ház-ban, ház-aktól; asztalos-nak, asztalos-okhoz; néptanács-tól, néptanács-okkal etc. We applied the same methods also in the case of the words with roots which comprise letters with variable values.

4.1. In view of the complications of the flexional system of Hungarian verbs and nouns, we have inserted certain restrictions. In the course of declension we dealt only with the reduced paradigm which comprises 17, respectively 34 forms without taking into account certain enlargements which express different relations [5].

The processed material of the Hungarian parts from L. Hadrovics and L. Gáldi "Hungarian-Russian Dictionary" [4] complemented with the "Explanatory Dictionary of Hungarian Language" [6] led us to determine the fact that the first 1400 nouns fall into 72 types. The great number of types is also the result of the law of vocalic harmony.

4.2. We established the conjugation types on the basis of a reduced system which contains only the indicative and conditional mood of the Hungarian verb. The future indicative was not taken into account, because it is formed with the auxiliary verb fog + the infinitive of the conjugated verb. The practical needs of MT obliged us to consider for this category a more adequate syntactical treatment than the morphological one [7].

From the point of view of MT a rather difficult problem is presented by the verbs with separable prefix. The treatment of the verb class with the ending -ik in the 3rd person singular also posed an uncommon problem. As this flexional system is decomposing, the cases of interpenetration of the paradigms are frequent, for example: ábrándozok - ábrándozom; ábrándoz - ábrándozik etc., which makes difficult the systematization of this class of verbs.

In the course of classification we applied the formal principle also to the treatment of the critical vowel considered as a part of the ending. For example: ablakom, várom, tanitanék were segmented as the following: ablak-om , vár-om, tanit-anék etc.

4.2.1. On the basis of the capacity of the Hungarian verbs to be transitive and non-transitive, we classified them in the following 3 classes of conjugation types:

a) verbs which have only subjective conjugation,

b) verbs which have subjective as well as objective conjugation,

c) verbs which have only objective conjugation.

4.2.2. Within these classes we have established the types of conjugation taking into account the palatal-velar correlation which creates in Hungarian a

specific relationship among the types of the various classes. A round number of 12000 verbs in Hungarian are distributed into 65 types of conjugation.

5.0. Although our investigations to classify the words of the three languages did not attain the final stage, it may be said that the language material is suited to a good formalization within some proper and comprehensive logical models. The discrete quantifiable nature of the large language material processed enables us to conclude that the application of the methods of the set-theory will give the possibility of its good algorithmization. The first tests of relational (syntactical) analysis also prove the possibility of expressing the algorithm in a language of orders and codes accesible to a computer with the capacity of rapid memory varying between 4-6 x 1024 and the external memory with a capacity of more ten thousand memory cells.

REFERENCES

[1] The authors would like to acknowledge the contributions of I. Birta, L. Chirvai, V. Drondoe, E. Hales, E. Kovács - Bölöni, H. Schuster, S. Soltyski, E. Székely and S. Székely.

[2] Dicţionarul limbii române literare contemporane, I-IV. Bucureşti, 1951.

[3] S.I. Ožegov: Slovar russkogo jazyka, 4th edition, Moscow, 1960.

[4] L. Hadrovics, L. Gáldi: Magyar-Orosz Szótár, Budapest, 1951. 995p.

[5] L. Antal: A magyar esetrendszer, Budapest, 1961.

[6] A Magyar Nyelv Értelmező Szótára I-VII. Budapest, 1959-1962.

[7] Otherwise L. Antal is of the same opinion in his work "On the Hungarian Verb". L. Antal: Gondolatok a magyar igéről, MNY., LVII. 3. 1961. 273-279p.

ИСПОЛЬЗОВАНИЕ ПЕРФОКАРТНЫХ МАШИН ПРИ МНОГОСТОРОННЕМ АНАЛИЗЕ ТЕКСТОВ

Ярмила Паневова

1. Сначала мы хотим коротко упомянуть о тех задачах, при решении которых были в нашей группе, которая работает при Карловом университете в Праге, использованы перфокартные машины и на основе которых у нас создался первый опыт в области механизации лингвистической работы:

а. В 1960 г. был составлен частотный словарь из математических текстов. Это был простой тип задачи, когда без предварительной лингвистической подготовки на каждую перфокарту было перенесено одно слово из текста. Алфавитное упорядочение и подсчет одинаковых словоформ - эта задача была решена машинами: сортировкой и табулятором. Чтобы создать инвентарь слов, встречающихся в английском тексте по электронике с целью составления словаря для машинного перевода с английского на чешский язык, мы обрабатывали с помощью перфокартных машин словарь объема в 100 000 словоформ. Текст сначала подвергается некоторой лингвистической обработка, т.е. к словоформе присоединяется характеристика, обозначающая ее часть речи, далее минимальный контекст /одно непосредственно последующее слово/ и чешский эквивалент данного слова. Предполагаем, что результатом машинной обработки получится не только в алфавитным порядке составленный словарь английских словоформ, но внутри омонимных словоформ получим упорядочение относительно частей речи и относительно чешского эквивалента. Эта внутренняя организация английского словаря будет служить исходным пунктом при поисках контекстуальных критериев для решения омонимии - лексикальной и грамматической.

б. При поисках оптимального способа хранения слов в памяти вычислительной машины проводилось исследование частотности чешских графем и их комбинаций в зависимости от позиции. Это проводилось тоже при помощи перфокартных машин.

в. На перфокарты переносились отдельные морфологические категории, выраженные формами чешских глаголов, взятых из сплошного текста. Исследовалась частотность отдельных категорий и их комбинаций /напр. было установлено, что чаще всего встречаются в чешском специальном тексте глаголы 3. лица ед. числа 4. спряжения /prosi/ , если мы имеем в виду именно комбинации категорий лицо, число, спряжение. Эти исследования были практически использованы при составлении алгоритмов морфологического анализа и синтеза.

г. Обзор разных типов работ мы закончим тем, что перфокартные машины были использованы при изучении обиходно-разговорного чешского языка. Отдельные типичные явления разговорной речи, характеризующие высказывание с точки зрения нормы литературного языка и отклонений от нее, переносятся на перфокарты и исследуется взаимосочетаемость явлений разных стилей в рамках одного высказывания вместе с данными о частотности.

Мы убедились, что использование перфокартных машин как средства механизации лингвистических работ выгодно там, где надо обработать большое количество материала и классифицировать его по разным критериям и именно там, где критерии классификации надо комбинировать - т.е. где количество и порядок параметров меняются.

Распределение перфокарты и данные, переносящиеся на перфокарты, можно установить так, чтобы слово из текста было однозначно характеризовано со всех точек зрения так, чтобы один комплект перфокарт служил основным ма-

териалом для исследования самых различных проблем из области чешской грамматики, в частности морфологии и синтаксиса. С этой точки зрения и с этой целью у нас проводился подробный анализ текстов, взятых из книг и статей по электронике и проанализированные слова со своими характеристиками по заданным категориям переносились на перфокарты.Категории, которые у каждого слова определялись и перфорировались, были установлены в связи с основной целью, которой является изучение грамматического строя чешского языка для подготовки анализа и синтеза для машинного перевода, для монографической обработки частных проблем в области чешской морфологии, синтаксиса и смысловой структуры предложения. Такие частные проблемы обрабатываются, учитывая возможность построения генеративной грамматики чешского языка с несколькими уровнями[1], и учивывая возможность использовать самый верхний план языка - план смысловой структуры предложения в роли языка-посредника[2].

2. Теперь мы хотим перейти к конкретному описанию данных и их переносу на перфокарты. Употребляются перфокарты, которые разделены в две половины - в каждой из них находится 45 колонн. Эти перфокарты обрабатываются буквенно-цифровыми машинами чехословацкой марки АРИТМА. В нижней половине перфокарты перфорируется слово из текста, далее 5 последних букв этого слова обратно /78-82 колонна/. Далее на перфокарте есть указание о том, где слово находится в тексте, т.е. его порядковый номер в предложении, порядковый номер предложения, страница и обозначение текста. Прежде всего нужно было установить, что считать единицей предложения, которой соответствует одна перфокарта: в связи с целью анализа единицей считается знаменительное слово /вместе со своими служебными словами/, Это значит, что на одной перфокарте находится предлог вместе с существительным, союз вместе со словом, которое им присоединяется, на одной перфокарте есть аналитическая форма глагола и т.д.Анализ проводится всегда внутри одного предложения. Контекстуальные связи пока не учитывались.

В морфологической части анализа устанавливаются кроме части речи /разбиение на части речи здесь более мелкое чем в школьных грамматиках, напр. местоимения разбиваются на несколько частей речи в зависимости от того, выступают ли они в предложении в роли сущ., прилаг. и т.д., в зависимости от типа склонения и в зависимости от признака относительности или отсутствия этого признака/ еще остальные формально-морфологические категории как падеж, число, род, тип склонения у имен; лицо, число, время, залог, вид,наклонение, тип спряжения у глаголов, причем при установлении этих категорий исходим всегда из формы. Это значит, что напр. время глагола совершенного вида в форме настоящего времени обозначается как настоящее, несмотря на временное значение этой формы, или форма превосходной степени прилагательного в значении элатива обозначается как суперлатив, несмотря на ее другое значение. Наряду с этим тут обозначено окончание и чередование в корне. Разбиение на типы склонения здесь также намного мельче, чем в традиционных грамматиках. Нужно, чтобы по типу склонения однозначно была определена каждая форма данного слова /напр. тип hrad разбивается на три подтипа в зависимости от окончания предложного падежа -e; -u или -e и -u/. Такой морфологический анализ требует конечно уже разработанных морфологических категорий, напр. заданного уже и проведенного списка типов склонения, типов чередований и т.п. В связи с тем, что у нас готовы почти полные алгоритмы анализа и синтеза чешской морфологии, можно было считать эти категории заданными.

Задавая синтаксические категории, по которым текст будет анализироваться, мы исходили из традиционного чешского синтаксиса, систематически описанного В. Шмилауэром[3] и сделали только несколько отклонений, связанных с нашей целью. У слова обозначено, в роли какого члена предложения оно выступает, далее порядковый номер его главного слова /по терминологии Н.Д. Андреева[4] : тектоглиф/, причем только предикат главного предложения имеет тектоглиф ноль. Придаточное предложение считается эквивалентом члена предложения в согласии с традицией чешского синтаксиса и предикат придаточного предложения несет информацию о типе предложения. Из области синтаксиса еще обозначается у слова, какой формой оно управляет в случае, если это предсказание в данном тексте выполнено, и далее - данные о богатстве и типе распространения данного слова. На перфокарте обозначена также часть речи главного слова, от которого данное слово зависит. Вообще мы стремились охватить как можно больше информаций у одного слова, значит на одной перфокарте, чтобы не было нужно часто прибегать к поискам контекста.

Далее мы считали важным установить основной репертуар категорий из верхнего семантического плана языка. В этот план не включена пока лексикальная семантика, поэтому более удобным считаем термин план смыслового строения предложения[5]. Это - часть, которая наименее устойчива и которая подвергается дальнейшей обработке, хотя и здесь можно использовать с успехом напр. синтаксис Шмилауэра, где приводятся семантические категории, выраженные морфологическими и синтаксическими средствами. В этой области устанавливаются, во первых, категории семантического синтаксиса как агенс действия, патиенс действия, действие само, детерминация. Далее определяется категория т.н. семантической части речи - здесь различается в основном категория субстанции, качества, процесса, обстоятельства. Субстанция напр. выражается, как правило, существительным, но она может перейти в некоторых случаях в прилагательное, если детерминирует другое существительное и выражает отношение между ними /напр. стекло х стеклянный предмет - это отношение двух существительных/. Или семантисеское прилагательное может перейти в наречие, если оно определяет глагол - выражает свойство, признак глагола - громкий х громко петь. Мы надеемся, что эти переходы частей речи - Куриловичу[6] синтаксические деривации - можно обработать внутри отдельных языков, так как отдельные языки сильно расходятся в возможностях этих деривaций, и что не следует их различать уже на семантическом уровне.

К этому плану языка мы относим тоже категории т.н. морфологической семантики. Это в сущности значения разных форм, которые объясняются в грамматиках иногда в разделе синтаксиса, иногда в специальном разделе о значении и опотреблении форм, но нам кажется, что они являются общими для языков, несмотря на разные средства выражения в разных языках. Это - категории такого рода, как у глаголов временное значение /для обозначения времени существуют в чешском языке три формы, но временных значений больше, существует еще вневременное значение, одновременность, преждевременность и т.п., иногда, как известно, форма выражения времени и временное значение не совпадают, напр. в предложении завтра я иду в кино/.

Далее, сюда входят семантические значения, выраженные разными формами имен, наречиями и разными типами придаточных предложений, как орудие действия, цель, причина, сравнение, время, место и т.д. Здесь мы считаем открытой возможность использовать этот материал для изучения вопросов лексической се-

мантики. В специальном тексте будет, наверное, замкнутая группа слов, выступающих в роли орудия действия, или группа временных слов. Интересно будет изучать слова, которые входят одновременно в несколько таких групп.

3. Приведем несколько замечаний насчет технической стороны переноса такого количества данных на перфокарты. Как уже сказано, грамматические категории перфорируются в верхней половине перфокарты. Так как мы работаем с буквенно-цифровыми машинами, мы выбираем символы, которыми обозначаются отдельные категории, как правило, алфавитные. Это выгодно там, где в данной категории /напр.: морфологическая семантика имен/ есть много элементов. При использовании алфавитных символов поместится в одном столбце 29 разных знаков в отличие от 10 цифровых в одном столбце. С другой стороны, использование алфавитной символики неэкономно, т.к. для сортирования алфавитных знаков машины АРИТМА требуют трех картоходов; после первого картохода рассортировано первых десять букв /первых десять в алфавитном порядке или десять самых частых букв алфавита/, остальные перфокарты идут второй раз через сортировку и сортируется вторая десятка букв и т.д. Десять цифровых знаков сортируются одним картоходом. Но нам кажется, что наряду с экономией, что касается количества картоходов, нужно учитывать также трудность, которую представляет работа с символами для человека, который готовит материал. С его точки зрения удобно, конечно, употреблять такие символы, которые легко запоминаются .

Учитывая ограниченный объем перфокарты, мы не укладываем каждую категорию на отдельный столбец, но устанавливаем комбинации взаимоисключающихся категорий, напр. одна колонна для падежа и число /т.с. особый символ для именит. пад. ед. числа, особый для им. мн. числа/. Дальше можно сэкономить столбцы тем, что в столбец, который у прилагательных несет информацию о степени сравнения, использован для другой информации у слов, не изменяющихся по степеням сравнения.

Эти подробные характеристики, приведенные на перфокартах, дают возможность различных и богатых комбинаций критериев классификации. Можно исследовать отношение между разными уровнями языка, включая данные о частотности. Так, можно проверить, во скольких случаях агенс действия является синтаксическим подлежащим предложения, во скольких случаях обстоятельством производителя действия и т.д.; как часто и при каких условиях окончание -u является окончанием родительного, дательного, предложного падежей; какие окончания сопровождаются теми или иными чередованиями и т.п. При сортировке материала можно поступать от формы к функции /как при анализе для машинного перевода/, или наоборот, можно исходить из категорий семантического уровня и дойти к характеристикам на уровне синтаксическом и т.д. Чтобы достичь этого, мы провели сортировку в нескольких этапах. На первом этапе исходилось из семантического члена предложения, далее сортировалось по семантической части речи, по данным морфологической семантики к формально-синтаксическому члену предложения и части речи. Результатом сортирования есть порядок перфокарт относительно символов в столбцах, по которым сортировка проводилась. Табулятор печатает данные с этих перфокарт и автоматически вырабатывает итоги. Благодаря тому, что табулятор способен выносить пять резрядов итогов, можем автоматически подсчитать частотность групп из пяти отдельных данных, по которым мы сортировали.

Пример:

Семантический член предложения	Семантическая часть речи	Морфологическая семантика		Член предложения		Часть речи	
D	V	Θ		A		A	540
D	V	Θ		A		S	42
D	V	Θ	901	A	901	V	319
D	V	T		A	7	S	7
D	V	T		D		S	50
D	V	T	63	D	56	V	6
D	V	•		•		•	
•	•	•		•		•	
I	II	III		IV		V	

Значение символов в рубриках

I	D	детерминация
II	V	глагол
III	Θ	беспризнаковая семантика /только субординация/
	T	временное значение
IV	A	аттрибут
	D	обстоятельство
V	A	прилагательное
	S	существительное
	V	глагол

На втором этапе нам интересно не только то, что, скажем время выражено существительным столько раз, но нужно знать, - в какой форме стоит существительное, с каким предлогом оно сочетается и т.д., поэтому на втором этапе доходим до формы слова. На третьем этапе мы хотели дойти до морфонологического облика слова поэтому сортировалось по морфологическим категориям - падеж, род, тип склонения, окончание и чередование, у глаголов вид, время, залог, наклонение, окончание и чередование. Таким образом доходим до графического оформления слова.

Возможности использования, надеемся, вытекают из сказанного. Из табулятором напечатанного порядка и итогов и из таблиц, которые на этой основе составляем, можно установить сразу некоторые интересные соотношения[7]. Но это, конечно, есть только основа для дальнейших исследований, когда необходимо искать в чешском языке условия, при каких выбирается за определенный семантический символ какая чешская форма /напр. форма придаточного предложения, форма номинализации, или есть возможность выбрать обе формы/.

Остается еще сказать, что мы таким образом обработали около 80 страниц электротехнического текста, которые представляют 22 000 перфокарт.

4. Возможностей использования перфокартных машин в лингвистике существует много. Но ориентируясь на использование перфокартных машин в лингвистике нельзя забывать о трудностях, которые с этим связаны. Нельзя полагаться на то, что ошибки при механической обработке исключены. Ошибки могут возникнуть уже при перфорировании, они исправляются после контроля, но не исключено, что на исправленной перфокарте возникают новые. Повторяющийся контроль замедляет работу. Даже сортировка и табулятор иногда допускают ошибки. Но мы уверены в том, что все-таки процент ошибок будет низший, чем процент ошибок,

допущенных человеком при классификации большого количества материала.

Хотим подчеркнуть, что использование перфокартных машин для нами описанной задачи не лишает лингвистов требовательной и трудной работы, связанной с анализом материала, установлением системы классификации и т.д. Но мы считаем преимуществом необходимость строгого установления критериев анализа, к которому нас заставляет использование средств механизации.

Кроме лингвистической подготовки с использованием перфокартных машин, связано много, менее квалифицированной, работы, трудной и требовательной, что касается времени. Мы встретились с трудностями того типа, что работники вычислительных центров редко работают с буквенными данными и что наши задачи для них сложнее чем те, которые они постоянно решают.

Вообще можно сказать, что при решении лингвистических задач, не требующих сложных операций, можно ориентироваться на использование перфокартных машин и в будущем, но у более сложных, требующих лингвистического понимания, будет более целесообразным ориентироваться на использование вычислительных машин с более оперативными средствами входа.

Л И Т Е Р А Т У Р А

1. P. Sgall, Zur Frage der Ebenen im Sprachsystem, Travaux linguistiques de Prague, Prague 1964, 95-106 p.
   P.Sgall, Ein mehrstufiges generatives System, Kybernetika 2, 1966, 181-190 p.
   P. Sgall, Generativni popis a česká deklinace, Praha, ČSAV (в печати).
2. P. Sgall, Převodni jazyk a teorie gramatiky, Slovo a slovesnost 24, 1963, 114-128 p.
   P. Sgall, The Intermediate Language in Machine Translation and the Theory of Grammar, Computational Linguistics II, Budapest, 1963, 35-62 p.
3. V. Šmilauer, Novočeská skladba, Praha, 1947.
4. N.D. Andrejev, Машинный перевод и проблема языка-посредника ВЯ 1957, 5, 117 сл.
5. M. Dokulil - F. Daneš, K tzv. veznamové a mluvnické stavbě věty, O vědeckém poznáni soudobych jazeku, Praha, 1958, 231-246 p.
6. J. Kurylowicz, Dérivation lexicale et dérivation syntaxique, Bulletin de la Soc. ling. de Paris, 37, 1936, 79-92 p.
7. Я. Паненова, Разбор электротехнических текстов, Prague Bulletin of Mathematical Linguistics, 4, 1966, 3-25 p.

PROBLEMS, PLANS ET POSSIBILITÉS ACTUELLES DE LA MÉCANISATION ET DE L'AUTOMATION DANS LA LINGUISTIQUE

Jitka Stindlová

1. Avant-propos
   Contact de la linguistique avec la technique et les sciences exactes
2. Méthodes appliquées au rassemblement et à l'organisation de la documentation linguistique auxiliaire; rangement par ordre alphabétique
3. Possibilités du traîtement mécano-graphique; méthodes opératoires propres à l'homme et à la machine
4. Formes d'implantation des caractéristiques dans les machines
5. Application des machines, avantageuse à l'heure actuelle, à certains domaines des travaux linguistiques:
   A. Mise au point des matériaux constitutifs, à savoir le rassemblement, la description, la conservation et le traîtement de la documentation
   B. Exploitation des ouvrages linguistiques déjà existants; enregistrements tirés des dictionnaires, grammaires, manuels, etc.
   C. Critique textuelle
6. Conclusions
   Les récentes possibilités de travail engendrent des tâches nouvelles
   Nécessité de coordonner les travaux et d'échanger les expériences

1. Avant-propos

Parmi les sciences sociales, la linguistique est une des premières à s'approprier les moyens opératoires efficaces qu'offre la technique de nos jours.

Pourquoi la linguistique en premier lieu? L'effort visant à la mettre en contact avec la technique et les sciences exactes, pour l'approcher de celles-ci, n'est pas, en effet, un souci qui préoccupe seulement les linguistes. D'une part, il est vrai que la linguistique en tant que science explorant les systèmes aussi complexes que les langues naturelles,a tout intérêt à profiter, pour ses propres fins,des moyens techniques récents; d'autre part la technique qui, dans l'effort d'étendre l'automation au domaine de la communication et du traîtement des informations formulées en langue soit écrite soit parlée, cherche à résoudre certains problèmes. Quelques uns de ces problèmes: la mise au point d'un système de lecture automatique susceptible de percevoir un enregistrement graphique, imprimé écrit à la main, l'analyse automatique de la langue parlée et la production de la langue syntétique, donc la tâche pratique de concevoir pour les machines l'entrée et la sortie parlées, la traduction automatique d'une langue en une autre langue soit directement, soit par l'entremise d'un intermédiaire, etc. Tout ceci nécessite aussi le concours de la linguistique et des linguistes. Parallèlement à ces tendances, menant l'une vers l'autre la linguistique et la technique, il se présente encore, avec l'apparition de la cybernétique, science des systèmes complexes, la possibilité d'échanger et de confronter les résultats et les méthodes d'étude de ces systèmes et du fonctionnement des unités qui les constituent. Or la linguistique qui, s'appliquant dans ces dernières dizaines d'années à saisir le système synchrone de la langue, sa structure, a réussi à obtenir des résultats concrets notamment dans l'étude du système phonologique, est en mesure de prêter son concours, au point de vue des méthodes, aux disciplines dont elle s'approche à présent et qui s'attachent à une caractérisation et à la discription du fonctionnement des systèmes complexes.

En cherchant à établir, entre la linguistique,la technique et les sciences exactes, un contact aussi intime et opératoire que possible, on commence à appliquer de nouvelles méthodes de description et de formulation exactes du phénomène examiné. L'usage est fait des descriptions et procédés mis au point par les sciences exactes, et en particulier par les mathématiques. Les paroles ne sont plus le seul instrument de la formulation, les symboles offrant des possibilités fort intéressantes; les symboles et les schématisations se substituent parfois avantageusement aux longues formulations verbales, souvent peu exactes.

Les relations nouées récemment entre la technique et la linguistique sont importantes pour cette dernière en premier lieu dans le domaine des méthodes . La technique implique l'exactitude et impose un travail fort détaillé, tout en permettant d'achever une abstraction poussée. C'est grâce à elle seulement que nous, linguistes, arrivons dans bien des domaines à l'application des méthodes d'un niveau supérieur.

Je n'ai pas la possibilité et je ne me propose d'ailleurs pas de consacrer mon rapport à la totalité des tâches délicates qui se posent dans la lin-

guistique et incombent aux linguistes, aux problèmes qui s'y rattachent et aux réponses respectives données dans le monde et dans notre pays. Je me borne à la question d'application de la technique à la linguistique. Je tâcherai de mettre au jour, par exemple, les points où la technique moderne est en mesure de faciliter nos travaux et les moyens techniques correspondant à cette fin et dont disposerais sous peu dans nos études.

2. Méthodes appliquées au rassemblement et à l'organisation de la documentation linguistique auxiliaire; rangement par ordre alphabétique

Dans l'exploration des systèmes complexes tels que les langues naturelles, exploration portant sur le système lui-même aussi bien que sur son fonctionnement sous forme de phénomènes linguistiques concrets, la linguistique doit se baser sur une multitude d'observations sans lesquelles elle ne saurait ni expérimenter, ni prédire purement et simplement; aussi a-t-elle à rassembler constamment, pour ses propres fins, une vaste documentation sous forme d'inventaires spéciaux à fiches de citations. Ces inventaires, sorte d'archives d'étude, ont pour fonction de justifier chaque phénomène linguistique à soumettre à un examen objectif; c'est à dire que tout phénomène étudié doit être représenté par un nombre satisfaisant de fiches, permettant de définir sa forme et son contenu ainsi que de généraliser sa fonction. Les inventaires organisés jusqu'à présent ne se prêtaient pas, bien entendu, à une application de plusieurs points de vue à la fois, toute étude individuelle entreprise sous un aspect différent imposant soit l'établissement d'un inventaire spécial, soit l'extraction de l'inventaire général de fichiers spécialisés. Les regroupements ad hoc d'un tel inventaire sont pratiquement irréalisables. A l'heure actuelle, il est inconcevable d'établir des fichiers spécialisés, en partant d'immenses inventaires qui ont été organisés aux fins de la composition de dictionnaires. Ainsi, l'idée d'employer la carte perforée en qualité de fiche "mobile" est due justement au besoin d'exploiter tout ce qui se cache dans une documentation linguistique, mais que son organisation traditionelle, n'envisageant qu'un seul but d'utilisation, rend inutilisable et inaccessible.

Prenons à titre d'exemple l'inventaire lexicographique de l'Institut de la Langue tchèque, comprenant à présent quelques dix millions de fiches. En séparer systématiquement une partie, pour justifier un certain phénomène linguistique par exemple les verbes, imposerait un travail énorme, sans parler de l'atteinte qui serait ainsi portée, pour une longue durée, au système intégral de l'inventaire. Et encore une séparation de toutes les fiches relatives aux verbes ne serait-elle probablement que peu féconde, à moins d'être suivie d'un rangement systématique spécial; sans spécialisation, cette immense documentation ne ferait, en effet, qu'accabler ceux qui auraient à l'exploiter. L'inventaire lexicographique de la langue tchèque contemporaine fut fondé en 1911. Pendant longtemps, il ne constitua pas un seul système alphabétique, car il se composait de nombreux fichiers indépendants, au nombre de 2.200 environ. Chacun d'entre eux représentait une documentation autonome, relative au vocabulaire d'un de nos écrivains, à une certaine période dans le journalisme ou la littérature technique par exemple, aux ouvrages d'une époque donnée. Même

au moment de la rédaction du premier volume du Dictionnaire manuel de la langue tchèque en 1936, ces deux milliers de fichiers n'avaient pas encore été réunis, mais se présentaient sous forme de 36 systèmes alphabétiques variés. Au cours de la rédaction du premier volume du dictionnaire, il s'avéra impossible, bien entendu, d'utiliser ce grand nombre de fichiers, bien qu'il fût déjà restreint; aussi réunit-on la documentation entière en un seul système alphabétique. Mais il en résulta en même temps la disparition des inventaires spécialisés, prévus pour les travaux lexicologiques et lexicographiques des vocabulaires d'auteurs, pour les études stylistiques de différents écrivains et époques.

Pour se rendre compte des possibilités qu'offre à la linguistique l'emploi des machines, supposons que la documentation lexicographique dont dispose l'Institut soit transposée sur les cartes perforées. On n'aurait plus à se soucier du rangement pénible des fiches ni de l'ordre alphabétique dans l'inventaire établi; ces travaux seraient exécutés par les machines. En même temps, il se présenterait à tout moment la possibilité de séparer, de ce système alphabétique unique, toutes les fiches relatives aux ouvrages de n'importe quel auteur, Nemcová, Jirásek, Neruda, et de mettre ainsi au point une documentation spécialisée, permettant de rédiger des vocabulaires, d'auteurs.

Il va sans dire que les machines permettraient non seulement l'organisation rapide d'une documentation complète concernant un auteur donné, mais aussi celle d'un fichier relatif à n'importe lequel de ses ouvrages, possibilités d'un intérêt considérable pour la confrontation avec les éditions nouvelles et éditions critiques.

Dans le domaine de l'histoire de la littérature et de l'étude du style, il serait possible de prévoir, par exemple, l'établissement de vocabulaires utilisés par les auteurs d'une même origine locale ou sociale, traîtant des sujets analogues, s'appliquant au même genre, etc.

On pourrait songer à regrouper l'inventaire en fonction de l'évolution historique, alternative fort intéressante pour l'étude de l'histoire et de la culture. Avec l'emploi des machines, une documentation prévue pour la composition des dictionnaires permettrait même de déterminer à quel moment apparaît, pour la première fois, un terme ou un autre, dans quelle acception et à quelle date se stabilise une acception donnée.

Les machines n'auraient pas de difficultés à grouper la documentation selon les catégories de sources: termes pris dans la poésie, dans la prose, dans les belles-lettres en général, dans les journaux, dans les ouvrages scientifiques ou dans le domaine de la propagation des sciences. Il serait possible de séparer, par exemple, les termes techniques et spéciaux, pour les grouper selon les branches professionelles et vérifier ainsi la représentation de celles-ci dans l'inventaire. L'intérêt d'un tel groupement pour la mise au point d'une documentation lexicologique et la caractérisation des styles est indiscutable.

Le traîtement monographique de l'inventaire lexicologique permettrait également d'établir une documentation prévue pour l'étude de la grammaire et d'organiser des fichiers selon les parties du discours et leurs formes, selon les catégories et les caractéristiques grammaticales, y compris les données

relatives à la fréquence qui ne seraient guère réalisables sans l'utilisation des machines.

Ce ne sont là que des suggestions qu'il n'est plus possible d'accomplir à l'inverse, à l'aide des inventaires déjà existants. Le fait est que, malheureusement, ces alternatives d'emploi ne sont point des possibilités réelles, le rangement statique par ordre alphabétique de l'inventaire visant un seul but, interdisant de les mettre en oeuvre. Les inventaires rangés par ordre alphabétique ne donnent que la possibilité de trouver, à la place prévue, le terme cherché et de ramasser alphabétiquement les fiches destinées à la composition des dictionnaires. Le système alphabétique empêche toute utilisation de la documentation péniblement mise au point pour d'autres études spéciales, notamment pour celles de la grammaire et des styles.

Si la manipulation de l'inventaire se fait à la main, le systéme alphabétique offre d'importants avantages sur les rangements systématiques et il est, pour cette raison, pratiquement le seul à être utilisé. Il est d'un accés facile et il n'y a personne qui ne sache le manier. Le fichier organisé avec l'application du principe alphabétique est aisément tenu à jour. Les rangements systématiques sont toujours d'un caractère documentaire, le plus souvent d'un intérêt temporaire; leur entretien impose inévitablement un travail pénible.

3. Possibilités du traîtement mécano-graphique; méthodes opératoires propres à l'homme et à la machine

Seules les machines permettent l'application du principe alphabétique parallèlement aux organisations systématiques. Cette possibilité est due au fait que les documentations traîtées par machines n'ont point un caractére statique, les fiches pouvant être exploitées dans des buts mutiples, sans imposer un entretien pénible et difficile.

A partir d'un inventaire unique, comprenant toutes les informations complexes implantées sur cartes, les machines à perforer sont susceptibles d'exécuter le triage sous les aspects les plus variés, mettant au point une variété de fichiers spécifiques et réétablissant, après l'emploi de ceux-ci, l'inventaire initial, prêt à accomplir les tâches suivantes. Les machines permettent une manipulation rapide et aisée de la documentation, par rangement alphabétique aussi bien que systématique, et par des confrontations et combinaisons respectant tous les points de vue voulus. Cette possibilité de combiner et de confronter les principes les plus variés constitue l'avantage essentiel de cette méthode de traîtement, car les machines sont susceptibles d'accomplir les objectifs futurs qui ne se fone pas même sentir à l'heure actuelle, mais qui exigeront en premier lieu de nouvelles combinaisons imprévues des critères linguistiques et des points de vue de triage. L'organisation traditionelle n'offrait point une telle utilisation. Elle n'était pas moins coûteuse, mais le travail lié à la documentation prévue pour un seul but d'emploi, s'opposait à l'exploitation, n'offrant à personne la possibilité de mettre à profit, d'une manière rapide et aisée, les informations cachées dans l'inventaire. Ainsi on n'exagère point en disant que la carte perforée est un instrument créé, pour ainsi dire, à notre usage. Même en concurrence avec les moyens plus

rapides et plus mobiles, destinés à faire entrer l'information codifiée dans la machine, tels que le ruban magnétique, la carte perforée est loin de perdre sa fonction, car elle permet de combiner les informations codifiées avec celles qui ne le sont pas. La possibilité d'établir un nombre illimité de cartes revient à prévoir en même temps une mémoire sans limites de la machine. En sa qualité de fiche de documentation faisant partie de l'inventaire linguistique, la carte perforée est très avantageuse pour la conservation des informations, la science du language nécessitant toujours une multitude de faits et la carte étant susceptible de recevoir aussi, à l'exception des informations codifiées, les données de forme traditionnelle, comme les definitions, citations, notes et autres. L'information implantée sous forme de code sur la carte peut être lue et transformée, s'il y a lieu, pour être enscrite sur le ruban magnétique, etc., la transformation d'un code à l'autre n'étant qu'un problème d'ordre technique, conditionné par l'emploi d'un transformateur de code adéquat.

Les machines à perforer en combinaison avec les calculateurs sont susceptibles d'adapter la documentation à l'étude des rapports complexes, caractérisant toute langue, et de fournir en même temps les données spécifiant les relations d'ordre quantitatif, valables dans le domaine des rapports complexes.

Préalablement à la préparation du travail à exécuter par les machines en vue d'obtenir les résultats voulus, il importe toutefois de se familiariser avec la méthode opératoire propre à la machine et de la confronter avec l'activité de l'homme. La machine est en mesure de s'approprier des systèmes fort complexes, mais elle aborde et saisit le système d'une manière qui diffère de celle appliquée par l'homme. Cela revient à dire qu'à la méthode opératoire propre à l'homme, la machine en ajoute une autre, supplémentaire. Le travail de l'homme est synthétique et évolue en abréviations, le fonctionnement de la machine utilise des éléments et procède sous forme de pas isolés qui sont les différents traits caractéristiques et leurs combinaisons, constituant et séparant l'unité du systéme, aussi bien que les phases de la réalisation et le fonctionnement des unités de système. Aussi la machine permet-elle à l'homme de mettre en oeuvre les éléments caractéristiques et toutes leurs combinaisons théoriquement possibles, les différentes phases isolées de la réalisation des unités dans le système. L'homme a besoin de telles "sections", sous forme d'éléments et de phases, pour l'étude du système, mais il ne saurait pas les produire lui-même ou les tirer rapidement des fiches maniées à la main. Les inventaires rangés par ordre systématique, sous forme d'usage courant, sont plus ou moins de telles "sections". Il ne faut pas oublier cependant que leur conception est statique, et ne donne des renseignements que sur une section unique. L'organisation de ces inventaires est pénible et, une fois exploités, ils ne servent plus à rien. Qu'il nous soit permis d'ajouter, à titre d'exemple, quelques considérations sommaires sur le système morphologique et son exploitation par l'homme et par la machine.

Pour s'emparer du système et le retenir dans sa mémoire, l'homme se sert de types, représentant les caractéristiques du système qui ne refusent, sans nuances, que l'essentiel des types les plus productifs. Ceci fait, il est encore obligé de faire entrer en ligne de compte les écarts d'une fréquence im-

portante; notons dans cet ordre d'idées qu'en particulier les termes bien fréquents du fonds de base échappent souvent à la description synchrone du système.

Les types de cette catégorie sont inutilisables sur la machine qui caractérise sans ambiguité et avec toutes les nuances les types aussi bien productifs qu'improductifs, ces derniers pouvant être d'une fréquence importante dans la parole. Pour la machine, chaque unité, chaque phénomène doit être nettement déterminé et défini par la combinaison des éléments distinctifs, c'est-à-dire des traits caractérisant le phénomène. Pour parler en termes empruntés à la phonologie, on songe à une espèce de point d'intersection des symptômes, en leur qualité d'éléments constitutifs de l'unité de système. C'est dire que les machines travaillent, sont susceptibles de travailler et sont obligées de travailler avec les éléments distinctifs, avec les traits composant la caractéristique, ou mieux le caractère de l'unité de système, ainsi qu'avac les points d'intersection de ces éléments.

Alors que l'homme travaille avec les représentants des caractéristiques dans lesquels les éléments distinctifs se présentent sous une forme typisée et cumulée, et sont complètement supprimés comme éléments, la machine exige des éléments aux traits distinctifs et leurs points d'intersection.

Mais si le travail de la machine impose la mise en oeuvre des éléments, des traits distinctifs, cela ne veut point dire que ceux-ci sont indispensables pour chaque phase de son fonctionnement. Les combinaisons concrètes des éléments, leurs points d'intersection, engendrent des unités qui font office de modèles représentant les éléments distinctifs, une fois les unités d'un degré supérieur mises au point. Toutefois, au départ du traitement des unités de système, la machine n'est susceptible de travailler que d'une manière analytique, en employant les éléments. La définition correcte des éléments, des traits distinctifs et l'adoption d'un nombre adéquat de ces éléments, conditionnent la capacité de la machine de percevoir et de traiter, d'analyser et de synthétiser les unités, comme les machines définissent les faits (phénomènes) linguistiques sous forme de combinaisons d'éléments distinctifs, de points de vue de triage.

Ce système d'enregistrement permet même à la machine de traiter des combinaisons d'éléments distinctifs qui ne trouvent aucune réalisation dans la langue, ou ne l'ont pas trouvée jusqu'à présent, mais qui existent tout de même en théorie et sous une forme potentielle. C'est justement <u>l'éxistence théorique de toutes les combinaisons possibles dans lesquelles peuvent entrer les éléments (traits) distinctifs,</u> consituant et caractérisant un certain phénomène, <u>qui conditionne tout travail exact</u>. La possibilité de séparer n'importe lequel des points d'intersection théoriquement possibles, des traits distinctifs et de leurs combinaisons, garantit la validité permanente de la documentation. Aussi, cette possibilité fait-elle disparaître l'objection souvent formulée, à savoir qu'aucune préparation, aussi pénible qu'elle soit, ne permet de prévoir à l'heure actuelle ce qu'il faudra soumettre à l'étude dans l'avenir, de sorte qu'avant de donner à la documentation la forme imposée par les machines, la nécessité se présentera d'aborder celle-ci d'une manière différant des tâches que la science du language se propose de résoudre actuellement.

Les machines sont en mesure d'exécuter ce que l'homme ne saurait réaliser, c'est-à-dire de conserver les éléments dans la mémoire. Du point de vue du système synchrone, les éléments des caractéristiques qui le constituent ne varient guère, les changements ayant lieu dans le domaine de leurs combinaisons et de leur fonctionnement; aussi faudra-t-il, pour les faits individuels, procéder de temps à autre à des corrections. Au programme des éléments qui varient dans les situations identiques, s'éliminant les uns les autres dans le même rang, et des éléments qui entrent en combinaisons mutuelles dans des rangs différents, la machine saura toujours donnes une réponse prompte.

4. Formes d'implantation des caractéristiques dans les machines

Pour que la machine puisse s'approprier les informations, il faut les lui présenter sous une forme adéquate, c'est-à-dire dans le langage qui lui est propre, en code. Laissons de côté, pour le moment, l'aspect technique des informations, des codes, pour y revenir, s'il y a lieu, dans les discussions respectives.

Les caractéristiques, les points de vue du triage, implantés sur la carte, font partie d'une certaine clé qui peut être numérique ou, plus rarement, alphabétique. C'est-à-dire qu'il faut attribuer un certain chiffre, ou une certain lettre suivant le cas, aque information, à chaque point de vue de triage décisif pour le traitement de la documentation, et que pour le chiffre adopté il faut prévoir une certaine position sur la carte perforée. La combinaison du chiffre lui-même et de la place qui lui revient sur la carte, définit le point de vue de triage envisagé.

Nous entendons distinguer nettement la clé du code. Sour clé nous entendons le système des chiffres ou des lettres, symbolisant et distinguant en même temps les différentes idées, valeurs, données et caractéristiques. En employant une clé pour les enregistrements, c'est-à-dire certains chiffres pour symboliser certaines données, le code est le procédé technique employé pour implanter la clé adoptée. Chaque point de vue de triage, par exemple chaque trait distinctif, trouve son expression dans un certain chiffre auquel une position sur la carte perforée est obligatoirement attribuée. Dans la position prévue, le chiffre est traduit par le code qui est par exemple la perforation exécutée sur la carte. Le nombre des points de vue de triage, implantés dans la mémoire de la machine sous forme de perforations de la carte, peut être fort supérieur à celui des places réservées aux perforations; il peut atteindre des millions d'alternatives.

La réunion de plusieurs colonnes prévues sur la carte à perforer, une colonne présentant la possibilité d'employer les chiffres de 0 à 9, permet d'obtenir un champ de perforation; les positions des chiffres dans différentes colonnes représentent différentes valeurs, exprimant les unités, les dizaines, les centaines, etc. Cette conception multiplie sensiblement le nombre d'enregistrements possibles, traduisant les informations à communiquer à la machine. Dans le champ d'une carte peut être placée toujours une information d'un certain caractère. Dans un champ à deux colonnes où les chiffres implantés dans la première ou la deuxième d'entre elles représentent respectivement les dizaines et les unités les informations possibles sont au nombre de 100, au lieu de 2 x 10 dans deux colonnes indépendantes.

Un champ à trois colonnes offre la possibilité de 1000 informations au lieu de 3 x 10, donc 30 dans trois colonnes indépendantes, et la réunion dans un champ de 6 colonnes permet d'enregistrer sur chaque carte un cas sur un million de cas possibles, alors que 6 colonnes indépendantes n'offrent que l'implantation de 6 informations variées sur 60. Cela veut dire qu'une clé à six positions, un champ à six colonnes sur la carte à perforer, permet d'implanter sur les cartes une documentation des faits établie sous forme de classification décimale à six places.

5. Application des machines à certains domaines des travaux linguistiques

A l'heure actuelle, l'utilisation des machines s'offre essentiellement aux trois domaines suivants des travaux linguistiques:

A. Le premier est la mise au point des matériaux constitutifs, à savoir le rassemblement, la description, la conservation et le traitement de la documentation, et cela pour résoudre les problèmes qui avaient été examinés, jusqu'à présent, à l'aide d'inventaires statistiques organisés, dans la majorité des cas sur le principe de l'ordre alphabétique, de même que pour les problèmes qui se posent nouvellement à la linguistique et qui ne sauraient être résolus, facilement et sûrement, qu'au moyen de machines, à savoir les recherches d'ordre quantitatif.

Le rassemblement a pour fonction, au fond, d'établir une documentation permettant la caractérisation, la généralisation. La majorité des inventaires linguistiques ne furent pas et ne sont pas rassemblés pour obtenir des données totales caractérisant les faits linguistiques, bien qu'une description numérique, statistique des phénomènes soit intéressante même en linguistique. Toutefois, le traitement manuel de la documentation linguistique n'avait pas pour objet, jusqu'à présent, un volume de fiches susceptibles de représenter les relations quantitatives. Aussi la description statistique ne jouait-elle qu'un rôle subordonné, documentaire. Le plus souvent, les linguistes ne se proposaient guère d'arriver à une expression statistique de la documentation. Quoiqu'il en soit, et il faut s'en rendre compte, toute documentation, même celle qui n'a pas pour fonction de traduire certains faits en chiffres - à la différence par exemple des fichiers prévus pour la composition des dictionnaires d'acceptions - représente tout de même un certain phénomène sous forme de cas choisis dans un groupe donné et possède ainsi un caractère statistique.

Toute expression statistique d'ordre numérique d'un certain phénomène est conditionnée par un rassemblement préalable des données respectives, par une mise en évidence de la fréquence du phénomène examiné et par la détermination de ses combinaisons avec d'autres phénomènes. Or le rassemblement de la documentation est justement cette première phase de la généralisation d'ordre statistique.

Le fait même que toute documentation possède certains traits de caractère statistique, qu'elle est la première étape des travaux statistiques, est à la base de la possibilité d'employer, pour la mise au point de la documentation, des moyens techniques modernes disponibles aux statistiques, à savoir les machines à perforer et les calculateurs.

Les machines permettent d'organiser la documentation même dans les domaines caractérisés par un mouvement permanent, où un rassemblement statique

des documents n'est plus satisfaisant. Parmi ces domaines se range, par exemple, la documentation ayant pour objet des termes techniques et spéciaux. Il est inconcevable de conserver, sous forme d'une documentation facilement maniable, la terminologie de différentes branches sans avoir recours aux moyens techniques récents. Seules les machines permettent de trouver la solution du problème qui consiste à conserver la documentation respective sous une forme vraiment opératoire, permettant de la regrouper à volonté et de l'exploiter, de cette façon, à des fins théoriques et pratiques sur le plan des acceptions aussi bien que sur celui de l'expression formelle; les détails respectifs sont spécifiés plus bas. Dans notre cas de l'inventaire relatif aux dialectes, pour effectuer le rassemblement et l'organisation des documents prévus pour l'établissement d'un atlas des dialectes slaves, pour effectuer l'inventaire du langage parlé, comme dans tous les cas mentionnés, il est nécessaire de posséder des fiches de caractère dynamique, susceptibles de confrontation.

Les machines permettent d'obtenir et de traiter, rapidement et avec précision, une documentation complète des textes, des auteurs, et d'enregistrer tous les termes que les textes contiennent; notons que d'importantes expériences ont été faites à ce sujet au centre lexicographique de Besançon en France, sous la direction du professeur B. Quemada. En dehors des index de référence aux endroits respectifs des ouvrages examinés sous forme de termes rangés par ordre alphabétique, les machines offrent également la possibilité d'établir des fiches de concordance avec contextes, permettant d'aborder le traitement lexicographique, l'analyse sémantique et celle du style et, avant tout, d'une manière intégralement automatique, d'effectuer des recherches d'ordre quantitatif.

Les éléments de documentation traités par la méthode mécano-graphique, peuvent être examinés sous les critères de la grammaire. du sens et du style. Pour ces études, il est possible d'adopter le rangement par ordre alphabétique mais aussi à l'inverse, de commencer par la fin.

Le rangement inverse des mots (mots-formes) sous les formes qu'ils possèdent dans les textes, combiné avec l'enregistrement de leur fréquence, est d'une grande importance pour les études morphologiques. Le rangement inverse des formes de base, infinitifs, nominatifs, adjectifs au nominatif du masculin, est intéressant avant tout pour l'étude de la formation des mots. Pour obtenir des données sur la méthode à suivre, nous avons fait, à titre d'expérience, l'essai de traiter sous forme mécano-graphique un fichier total du texte d'un ouvrage de l'écrivain tchèque, K. Havlicek Borovsky, intitulé "Tyrolské elegie".

Le volume important de la documentation n'est pas un obstacle pour le travail des machines. Au contraire, l'emploi de celles-ci n'est rentable que s'il s'agit d'une quantité importante de documents à traiter; par rentabilité nous n'entendons pas, dans ce cas, les frais d'ordre matériel, mais les dépenses affectées à l'établissement du projet, du programme et de la mise au point de la documentation, préalablement au traitement.

Aussi est-il bien avantageux d'utiliser les machines pour l'étude d'une documentation totale des textes. Le programme est relativement simple et la documentation abondante. Au texte implanté par voie mécano-graphique peut se

joindre l'examen quantitatif; à cet effet, nous avons aussi préparé les documents, pris dans l'inventaire total. L'importance de l'étude quantitative, portant sur le système de la langue et en premier lieu sur le fonctionnement de celle-ci dans les paroles, est indiscutable. Les données précises,résultant du traitement statistique, donnent non seulement la possibilité d'aborder la généralisation linguistique, sous l'aspect synchrone, historique ou comparatif, mais elles servent en même temps à d'autres études, et sont indispensables pour la solution technique des problèmes qui se rattachent au transfert et au traitement des informations. Il importe de se concentrer de préférence, à l'heure actuelle, sur les questions signalées, à savoir sur l'élaboration le plus rapide possible d'une documentation nécessaire pour étudier les problèmes techniques, sur celle de la conception et des programmes des machines à traiter les informations. Sous réserve que la conception et les programmes à établir auront pour base une analyse quantitative poussée, les travaux exécutés par les moyens techniques utilisés seront plus efficaces et plus sûrs, tout en nous offrant, pour nos propres buts, des résultats précis.

B. Les machines permettent non seulement de procéder à la mécanisation du rassemblement, de la description, de la conservation et du traitement de la documentation, c'est-à-dire à la mise au point des matériaux conditionnant l'analyse et la généralisation scientifiques, mais elle nous offrent à la fois la possibilité d'exploiter les ouvrages linguistiques tout faits. Elles permettent d'utiliser les résultats conservés dans ces ouvrages aux fins de nouveaux travaux théoriques et pratiques, de vérifier et de perfectionner, s'il y a lieu, l'ouvrage examiné.

Dans la première étape de cette utilisation des machines, je songe à l'enregistrement par voie mécano-graphique du Dictionnaire de la langue tchèque littéraire, qui doit paraître sous forme de cahiers; deux volumes sont déjà sortis, le troisième est sous presse et le quatrième en rédaction. L'ouvrage doit être achevé au début de 1967. L'enregistrement envisagé nous permettra de donner aux travaux de rédaction toute précision voulue, dans le domaine de l'unification, de la révision, des références, etc. Nous mettrons ces résultats à profit lors de la révision de la première rédaction du dictionnaire et, en particulier, dans la deuxieme édition. Nous avons déjà essayé, à titre d'expérience, d'implanter sur les cartes perforées les mots compris sous la lettre "C" du Dictionnaire; les spécimens sont à voir aux cours des discussions. Nous aurons à notre disposition, en même temps, une vaste documentation pour les études lexicologiques, grammaticales et stylistiques, de caractère aussi bien pratique que théorique. Nous en profiterons, avant tout, pour l'étude du système lexicologique, c'est-à-dire du système qui est, à présent, plutôt prévu que défini et saisi. Le système comprenant une multitude d'unités, il est fort difficile de l'étudier sous une forme traditionnelle. A la meilleure connaissance du système contribuera sans doute également l'étude des rélations quantitatives respectives, s'appuyant sur le dictionnaire. Le traitement mécano-graphique du dictionnaire pourra servir, en outre, à d'autres objectifs encore qui ne sont qu'en voie de concrétisation à l'heure actuelle. La documentation qu'offre le dictionnaire, enregistrée sous la forme imposée par les machines respectives, peut servir de "base" à l'éta-

blissement d'un inventaire automatique; pour les détails respectifs, je renvoie aux discussions. La classification exacte des catégories de mots enregistrée par machines, portant également sur les caractéristiques morphologiques et morphonologiques, pourrait devenir la base du fonds des termes aux fins de la traduction automatique.

Dans l'étape suivante de l'exploitation des ouvrages existants, on abordera probablement un ouvrage relatif à la formation des mots, à rédiger collectivement par la section grammaire et style de l'Institut de la langue tchèque; le volume 1 - introduction théorique par M. Dokulil "Théorie de la dérivation des mots" (avec un résumé en anglais et russe) fut publié en 1963 (voir aussi le volume 2, consacré à la formation des substantifs, le volume 3, sur la formation des verbes et des adjectifs).

Il n'y aura pas de difficulté à traiter la documentation lexicologique des dictionnaires bilingues. Citons à titre d'exemple, qu'un inventaire précieux sans utilisation jusqu'à présent, est disponible, dans le dictionnaire allemand-tchèque de J. Dobrovsky et dans d'autres dictionnaires datant du dix-neuvième siècle.

Des travaux analogues furent réalisés par le professeur Quemada à Besançon qui élabora l'inversion du dictionnaire franco-hollandais (B. Quemada: L'exploitation mécanique des dictionnaires bilingues, Index français-flamand du Vocabulaire de Berlaiment, Bulletin d'Information IV. Besançon 1961 du Laboratoire d'analyse lexicologique).

On pourrait songer également à rendre accessibles et à accélérer les travaux ayant pour objet l'inventaire lexicologique à dix millions de fiches. A présent, l'étendue de la documentation constitue souvent un obstacle s'opposant à un traitement rapide, à l'établissement d'un aperçu maniable de la documentation et à la recherche facile des fiches respectives. Maints objectifs seraient réalisables, sans doute, par l'emploi des enregistrements mécano-graphiques, sous forme d'une espèce de "prise"; une prise sur la carte perforée pourrait contenir, par exemple, le mot et sa caractéristique grammaticale, la fréquence des documents où il se trouve et la fréquence de ses formes, l'enregistrement des ouvrages et auteurs où il se présente, l'information sur les sources de documentation et sur leur caractère, les dates des documents les plus récents et les plus anciens, la spécification de l'utilisation du mot dans le langage technique (pour quelles branches). Il serait possible d'indiquer également sur la carte si le mot est incorporé aux dictionnaires et les valeurs que ceux-ci lui attachent. Une caractérisation plus poussée des documents portant sur les différentes acceptions pourrait s'appuyer sur les acceptions et leur numérotage dans le Dictionnaire de la langue tchèque littéraire. A ces prises seraient à adjoindre, sous une forme analogue, les familles de mots et les groupes phraséologiques dans lesquels le terme apparaît. Un traitement pareil, à titre d'information, de l'inventaire sous forme de "prises" serait assez difficile à réaliser et il représenterait un travail plus ou moins égal à la première rédaction des fiches lexicographiques; mais il n'en est pas moins vrai qu'il serait utile de l'envisager.

C. Le troisième domaine d'une utilisation féconde des machines est celui de la critique textuelle. Les machines sont en mesure d'accomplir des colla-

tions et confrontations fort difficiles, elles peuvent fournir l'extrait des différences existant dans les textes et mettre à notre disposition les documents relatifs à la critique textuelle et aux travaux d'édition.

Nous ne faisons que mentionner cette tâche, plutôt à titre de possibilité future, comme les textes d'une certaine étendue exigent l'emploi non seulement des machines à perforer, mais à la fois d'un calculateur, indispensable pour les travaux de confrontation. On attend toujours l'établissement d'une conception adéquate d'un système de lecture, mettant au point la perception automatique du texte (imprimé ou écrit à la main). Avec un tel dispositif, il serait possible d'accélérer sensiblement les travaux respectifs. Toutefois, il n'y a plus de doute que ce dispositif sera d'une conception délicate imposant l'emploi d'un calculateur. Il en résulte ainsi que pour une certaine étendue de ces travaux, il sera impossible de s'en passer. Quoiqu'il en soit, nous avons l'intention de rassembler les expériences méthodiques relatives à la solution du problème en cause par l'utilisation des machines à perforer. Un des premiers vocabulaires d'auteurs organisés par voie mécano-graphique sera probablement l'inventaire lexicologique des ouvrages de l'écrivain tchèque, Petr Bezruc. Dans la premiere étape, les travaux se concentreront sur son recueil de poésies intitulé "Slezské pisne". Cet objectif nous demande en premier lieu d'étudier et de confronter toutes les éditions du recueil et de faire entrer en ligne de compte les interventions qui y ont été faites.

6. Conclusions

Les machines nous offrent ainsi la possibilité d'appliquer de nouvelles méthodes exactes aux travaux linguistiques, méthodes courantes dans les autres sciences notamment dans les mathématiques, et d'obtenir de cette façon des résultats précis, sûrs, complets et exempts d'opinions subjectives, d'approximations et de défauts dus au manque de documentation ou à l'aspect peu complexe.

Les nouvelles possibilités ainsi que les besoins du travail mécanique nous imposent toutefois d'accomplir une tâche fort délicate qui est de procéder à une analyse universelle et poussée des faits linguistiques, entreprise sous un aspect assez nouveau. Il faut arriver aux éléments caractéristiques et établir des clés adéquates, ainsi que définir les éléments distinctifs sur tous les plans; sur les plans supérieurs peut-être seulement théoriquement, en structure, sous forme de modèles. Ces préparatifs peuvent réussir s'ils bénéficient d'une large participation du public linguistique et technique. On peut prévoir que le traitement mécanique de la documentation linguistique deviendra progressivement la méthode opératoire plus ou moins générale; il est d'autant plus nécessaire d'adopter, de commun accord, des points de vue appropriés à la classification des différents domaines de la linguistique. Les moyens techniques disponibles offrent la possibilité d'incorporer ou de séparer les fichiers les plus différents et de reproduire aisément les inventaires disponibles. Ainsi, il sera possible de réunir ou de répartir, dans certains buts, les documentations variées. Or la combinaison de différentes documentations, par exemple en vue de composer un ouvrage lexicographique, ou de mettre au point un dictionnaire inverse, ou encore d'établir une image de la fréquence, est conditionnée par un accord préalable des clés pour un classement identique

des faits linguistiques d'importance essentielle. Lors de l'établissement des clés, il faut faire entrer en ligne de compte tous les objectifs et besoins existant dans différents domaines des travaux à exécuter. Ainsi, la classification fondamentale des mots, due au besoin de prendre les mots directement dans les textes respectifs, aura une forme différant de la forme consacrée pour les recherches traditionnelles, et encore une autre pour l'enregistrement des mots déjà incorporés au dictionnaire. Tous ces besoins sont à respecter dans la classification fondamentale qui sera ainsi leur synthèse.

Mon rapport n'a d'autre but que de contribuer aux discussions à venir, de donner une base dans certains domaines ou de présenter des suggestions dans d'autres. Aussi mon rapport est à prendre comme une sorte d'introduction aux discussions au cours desquelles je reviendrai avec plaisir à n'importe quelle question aussi bien qu'aux projets concrets.

M I S C E L L A N E A

ÜBER DIE EXPERIMENTE AN EINEM SPRACHSTATISTISCHEN AUTOMATEN

J. Kelemen

1. Auf dem Lehrstuhl für Fernmeldetechnik an der Technischen Universität zu Budapest wurde unter der Leitung von Prof. László Kozma ein sprachstatistischer Automat geplant, hergestellt und dem Sprachwissenschaftlichen Institut der Ungarischen Akademie der Wissenschaften im Jahre 1965. übergeben [1].

2. Seitdem haben wir verschiedene Experimente ausgeführt, um die Möglichkeiten und die Brauchbarkeit des Automaten bei sprachstatistischer Analyse ausgewählter schriftlicher Texte beobachten und beschreiben zu können.

Bisher haben wir literarische Texte analysiert: je 3 Novellen von zwei bekannten ungarischen Schriftstellern, D. Kosztolányi und Zs. Móricz; einzelne Teile aus dem soziographischen Werke von Gy. Illyés: A puszták népe d.h. Das Volk der ungarischen Puszta-Siedlungen (in Transdanubien); eine literaturgeschichtliche Abhandlung von J. Horváth und einen Teil aus einem sprachwissenschaftlichen Werke von G. Bárczi: A magyar nyelv életrajza (Biographie der ungarischen Sprache).

3. Die Gesichtspunkte der Textanalyse der Experimente waren: 1. Lautfrequenz (abgesehen von Assimilationserscheinungen), bzw. Buchstabenfrequenz (die "zusammengesetzten Buchstaben" als Zeichen eines bestimmten Phonems -cs, dz, dzs, gy, ly, ny, sz, ty, zs - diese als einzelne Buchstaben aufgefasst). 2. Frequenz der verschiedenen Silbentypen (Silbenqualitätsfrequenz); 3. Frequenz der Silbenlängetypen nach Laut-, bzw. Buchstabenzahl (ohne Rücksicht auf die Assimilationserscheinungen). 4. Frequenz der Wortlänge nach Laut- bzw. Buchstabenzahl (mit den unter Nr. 1. in diesem Absatz umschriebenen Einschränkungen). 5. Frequenz der Wortlänge nach Silbenzahl. 6. Wortartenfrequenz, bzw. Frequenz der Wortartenwerte in den gewählten Texten, mit Rücksicht auf die innere Einteilung der einzelnen Wortarten. 7. Frequenz der Satzlänge nach Wortbestand. 8. Frequenz der Aussage-, Frage- und Aufforderungs-/Ausrufungssätze, insofern sie sich in der Druckform der satzschliessenden Interpunktionszeichen spiegeln.

Man könnte natürlich auch andere Gesichtspunkte (wie Buchstabenverbindungen, morphologischen Aufbau der Wortformen, Gliederung und Aufbau der Sätze nach Syntagmen, Begriffskreise der Wörter) bei dem Automaten anwenden. Dazu müsste aber ein geeignetes Programm ausgearbeitet werden, die Kapazität des Automaten ist jedoch zur Analyse des Textes nach manchen Gesichtspunkten zu klein. Es lohnt sich auch nicht, so umfangreiche Analysen wie z.B. die der Laut-, bzw. Buchstabenverbindungen, durch diesen Automaten vorzunehmen, weil dazu mehrere Exemplare vom Lochband desselben Textes nötig wären, weil fehlerlose Vervielfältigung der Lochbänder eine mühsame, langsame Arbeit ist, und da man bei diesem Verfahren noch immer keine dazu nötige Dokumentation der einzelnen Laut-, bzw. Buchstabenverbindungen hätte. Für solche Arbeiten sind Lochkarten oder Magnetbänder bzw. Magnetplatten der Elektronenrechenzentren viel geeigneter.

4. Die Hauptbestandteile der Einrichtung der mechanisch-automatischen Textanalyse sind: ein Fernschreiber (Olivetti - Ivrea Telescriventi mit

5 Kanälen für Lochbandschrift), ein Abtaster, eine Programmtafel mit Einstellfeld und die Auswertstromkreise. Die beiden letzten Hauptbestandteile sind in einem Metallschrank angebracht.

5. Die einzelnen Arbeitsphasen der bisherigen Experimente sind: 1. Die Transkription (dem Wesen nach Transliteration) des Textes nach dem "Alphabet" des Automaten. 2. Kodierung des Textes mit Bezeichnung der Wortartenwerte der Wortformen. 3. Herstellung des Lochbandes mit Kodenummern. 4. Verbesserung des Lochbandes. 5. Programmierung des Automaten zur Textanalyse. 6. Erstes Abschreiben der Ziffern des Zählapparates des Automaten. 7. "Lesen" des kodierten Textes nach dem vorbereiteten Programm durch den Automaten. 8. Zweites Abschreiben der Ziffern des Zählapparates. 9. Subtraktion. 10. Tabellierung der Frequenzdaten des analysierten Textes. 11. Summierung der Frequenzdaten der verschiedenen Texte. 12. Sprachwissenschaftliche Bewertung der Frequenzdaten.

5.1. Zur Transkription (Transliteration) wird ein Olivetti-Fernschreiber (4) gebraucht. Bei dieser Phase wird der Text nur in Klarschrift geschrieben, ohne Lochen des Lochbandes.

5.1.1. Das internationale Alphabet wird folgendermassen modifiziert:

5.1.1.1. Das Lochen für q wird mit dem Lautwert des kurzen ö, das Lochen für w mit dem Lautwert des kurzen ü gebraucht. Wenn im Text ganz ausnahmsweise doch der Buchstabe q vorkommt, wird er in der Transkription durch die Buchstabenverbindung kv ersetzt und unter den Anmerkungen über die Arbeit aufgezeichnet; ebenso wird ein seltenes w des Textes durch vv ersetzt und aufgezeichnet.

5.1.1.2. Das Lochen für y (in Klarschrift erscheint es als ▼) hat eine dreifache Funktion: a) vor dem Buchstaben der Selbstlaute wird es als Zeichen der Vokallänge gebraucht (also ▼a bezeichnet á, ▼e steht für é usw.); b) vor dem ersten Buchstaben der Buchstabenverbindungen für einen Laut steht es anstatt des zweiten Elementes dieser Buchstabenverbindungen (z.B. ▼c bezeichnet cs = c; ▼d steht für dz usw.); c) wenn es vor dem Buchstaben h steht, bezeichnet die Verbindung ▼h den Laut ddz. Wenn das y ganz selten im Text mit anderem Wert vorkommt, wird es durch die Verbindung ij ersetzt und aufgezeichnet.

5.1.1.3. Das Lochen für x (in Klarschrift erscheint es als X) hat auch eine dreifache Funktion: a) vor dem ersten Buchstaben der Buchstabenverbindungen für einen Mitlaut steht es im allgemeinen teilweise anstatt des zweiten Elementes der Buchstabenverbindung, teilweise für die Bezeichnung der gekürzten Doppelung dieser Verbindungen, z.B. Xc steht für ccs, Xg für ggy usw.; b) die Verbindung Xd bezeichnet die kurze Affrikate dzs; c) die Verbindung Xh steht für langes (gekürzt verdoppeltes) ddzs. Wenn der Buchstabe x ganz selten im Text vorkommt, wird er durch die Verbindung kvs (=ksz) ersetzt und aufgezeichnet.

5.1.1.4. Das "Alphabet" für den Automaten sieht also folgendermassen aus:

Buchstabe oder Buchstabenverbindung		Buchstabe oder Buchstabenverbindung	
der Rechtschreibung	des Automaten	der Rechtschreibung	des Automaten
a	a	nny	Xn
á	va	o	o
b	b	ó	vo
c	c	ö	ö /Lochung für ursprüngliches q/
cs	vc		
ccs	Xc	ő	vö
d	d	p	p
dz	vd	/q	kv*/
ddz	vh	r	r
dzs	Xd	s	s
ddzs	Xh	sz	vs
e	e	ssz	Xs
é	ve	t	t
f	f	ty	vt
g	g	tty	Xt
gy	vg	u	u
ggy	Xg	u	vu
h	h	ü	ü /Lochung für ursprüngliches w/
i	i		
i	vi		
j	j	ü v /w	vü vv*/
k	k	/x	kvs*/
l	l	/y	ij*/
		z	z
ly	vl	zs	vz
lly	Xl	zzs	Xz
m	m		
n	n		
ny	vn		

5.1.1.5. Die Folgerichtigkeit der Transkription wird zwar durch die Bezeichnung der Laute ddz, dzs und ddzs gebrochen, diese kommen aber sehr selten vor, bedeuten also praktisch keine beträchtliche Schwierigkeit. Ebenso wird die Umschreibung für q, w, x, y des Textes durch kv, vv, ks, ij keine Schwierigkeit bei der Auswertung bedeuten, wenn man auf Grund der besonderen Aufzeichnungen bei Laut- bzw. Buchstaben- und Silbenstatistik diese wenigen Fälle berücksichtigt.

5.1.2. Die Transkription wurde so ausgebildet, dass jeder Buchstabe und jede Buchstabenverbindung nach einem Buchstabenwechsel geschrieben werden

kann und die Schrift binnen einer Silbe nicht durch Zeichenwechsel unterbrochen werden muss.

5.1.3. Die Silbengrenze wird durch das Zeichen / bezeichnet. Dieses Zeichen wird mit Anschlagen des Zeichenwechsels, die folgende Silbe mit Anschlagen des Buchstabenwechsels eingeleitet.

5.1.3.1. Anstatt Xc (für ccs) usw. wird in intervokaler Stellung "Buchstabe" v Buchstabe c Zeichenwechsel / Buchstabenwechsel "Buchstabe" v Buchstabe c gelocht (in der Klarschrift erscheint es als Vc/vc) usw., um die automatische Silbenzählung zu ermöglichen.

5.1.3.2. Assimilationserscheinungen, die sich in der Rechtschreibung nicht widerspiegeln, werden auch bei Silbentrennung nicht bezeichnet. Z.B. die Wortform bántsa wird nicht als bVan/vca, sondern als bvant/sa gelocht (natürlich mit der Lochung des Zeichen- und des Buchstabenwechsels, wo es notwendig ist). Die Silben werden also in Graphemverbindungen gelocht und analysiert.

5.1.3.3. Die Transkription der Assimilationserscheinungen und damit die Analyse der Silben als Phonemverbindungen wird erst später verwirklicht, da die Analyse sowohl der graphischen als auch der akustischen Seite des Textes dem Projekt nach auseinandergehalten werden soll.

5.1.3.4. Bei Silbengrenzen der Zusammensetzungen werden die einzelnen Gliedwörter als sprachlich selbständige Einheiten beurteilt, also Wortformen wie vasut als vas/vut (nicht als va/svut) in Silben geteilt.

5.1.4. Die Interpunktion wird mit wenigen Ausnahmen nach der internationalen Fernschreiberpraxis gelocht und geschrieben. Satzende wird immer durch einen Punkt bezeichnet, abgesehen davon, ob es sich um einen Aussagesatz, einen Ausrufungs-, Aufforderungssatz oder um einen Fragesatz handelt , um damit das automatische Satzzählen zu ermöglichen.

5.1.4.1. Fragesätze werden mit der Kombination "Punkt Fragezeichen" (.?) unterschieden.

5.1.4.2. Ausrufungs- und Aufforderungssätze, deren Satzende in der Rechtschreibung durch Ausrufungszeichen bezeichnet wird, werden mit der Verbindung von "Punkt Doppelpunkt" (.:) abgeschlossen.

5.1.4.3. Der Punkt in anderer Stellung (z.B. nach Abkürzungen) wird durch Gleichheitszeichen (=) ersetzt, um die Störungen des Satzzählens zu eliminieren.

5.1.4.4. Semikolon (;) wird in der Transkription durch die Verbindung "Punkt Strich" (.,) ersetzt.

5.1.5. Da bei den Experimenten auch auf den Wortartenwert der Wortformen Rücksicht genommen wurde, wurde nach den Wortformen auch eine symbolische dreistellige Zahl geschrieben, um Platz für die endgültige Bezeichnung des Wortartenwertes zu sichern und die Innervation der Bewegungen der endgültigen Lochbandschrift vorzubereiten.

5.1.6. Bei der Kontrolle der Transkription werden die eventuellen Schreibfehler an der Klarschrift verbessert.

5.2. Kodierung des Textes mit Bezeichnung der Wortartenwerte der Wortformen. Nach Verbesserung der Klarschrift trägt man eine dreistellige Zahl

als Symbol des Wortartenwertes nach der betreffenden Wortform ein. Das System der Wortartenwerte wurde so aufgestellt, dass die erste Ziffer von 1 bis 9 den Wert nach folgenden Wortarten bedeutet: 1 Zeitwort (Verbum); 2 Hauptwort (Substantivum); 3 Eigenschaftswort (Adjektivum); 4 Zahlwort (Numerale); 5 Fürwort (Pronomen); 6 Bestimmungswort (Adverb); 7 Verbalnomen bzw. Verbaladverb (Infinitiv; Partizip); 8 "Postposition"; 9 Artikel. Die anderen Wortarten werden durch eine zweistellige Zahl unterschieden, deren erste Ziffer immer Null (0) ist, also 01 (oder 02) Satzwort ("Interjektion"); 06 Adverbiales Präfix eines Verbstammes (ung. "igekötő"); 08 Bindewort (Konjunktion). Die zweite und dritte Ziffer nach 1 - 9 dient dazu, die Unterklasse oder Unterart der betreffenden Wortart, teilweise auch andere Unterschiede zu bezeichnen, die mit Wortartenwerten in Zusammenhang stehen. So bezeichnet z.B. 111 intransitives Zeitwort ohne Präfix, 118 dasselbe mit Präfix, 121 transitives Zeitwort mit Objekt ohne Präfix, 128 dasselbe mit Präfix usw.

5.3. Die Herstellung des Lochbandes geschieht durch einen Olivetti-Fernschreiber auf Grund des transkribierten Textes mit Kodierung der Wortartenwerte. Der Fernschreiber locht das 5-Kanäle-Lochband und schreibt gleichzeitig den Text in Klarschrift.

5.3.1. Wenn die Typistin bemerkt, dass ein Fehler gemacht wurde und in der Schrift vom Fehler nicht zu weit gelangt ist, kann der Fehler während des Schreibens verbessert werden. Den unteren, kleineren Arm des Lochapparaten stellt sie in die Stellung mit der Aufschrift ESCL, nachdem drückt sie den oberen, längeren Arm einmal, wenn sie um einen Schritt, zweimal, wenn sie um zwei Schritte den Fehler in der weiteren Schrift hinter sich gelassen hat usw. Dabei muss sie acht geben, auch den Gebrauch vom Wagenrücklauf, Zeilenvorschub, Buchstabenwechsel, Zeichenwechsel und Zwischenraum als einen selbständigen Schritt mitzuzählen. Nachdem bringt sie den unteren Arm in Stellung INCL und drückt den Buchstabenwechsel so viele Mal, wie viel Schritte mit dem oberen Arm das Lochband zurückgezogen wurde. Die Schrift wird von dem letzten fehlerlosen Buchstaben bzw. Zeichen ohne Zwischenraum fortgesetzt.

5.3.2. Die Wortform wird mit Silbengrenzen gelocht; die dreistellige Zahl wird ohne Zwischenraum nach der letzten Silbengrenze gelocht. Nach der Zahl des Wortartenwertes locht man einen Zwischenraum. Vor der ersten Wortform steht kein Zwischenraum; nach der Zahl des Wortartenwertes der letzten Wortform des Satzes locht man auch einen Zwischenraum und erst dann folgt die Lochung des satzschliessenden Interpunktonszeichens. So zählt der Automat durch die Lochung der Zwischenräume unmittelbar die Zahl der Wortformen des Textes.

5.4. Die Verbesserung des Lochbandes. Nach Beendigung der Herstellung des Lochbandes wird die Lochung durch Vergleich der gleichzeitig hergestellten Klarschrift mit dem Originaltext geprüft. Die Fehler werden in der Klarschrift bezeichnet, am Lochband herausgesucht und verbessert.

5.4.1. Wenn bei der Lochung nur ein Buchstabe oder ein Zeichen verfehlt wurde, wird das überflüssige Loch mit einem dünnen Klebeband zugeklebt, die weggebliebene Lochung mit einem Handlocher nachträglich vorgenommen.

5.4.2. Wenn eine Buchstaben- oder Zeichenlochung bzw. eine Buchstaben-

wechsel-, Zeichenwechsel-, Zwischenraum-, Zeilenvorschub oder Wagenrücklauflochung ausgeblieben ist, wenn irgendwo am Lochband mehrere Fehler entdeckt wurden, wird das Lochband nach der letzten fehlerlosen Lochung abgeschnitten, der richtige Textteil in ein neues Lochband fehlerlos gelocht, das ursprüngliche Lochband und das darunter gelegte fehlerlose neue zusammengeklebt. Wenn die Verbesserung nur aus wenigen (1 bis 2-3) Lochreihen besteht, wird das andere Stück des Originalbandes nicht unter, sondern über das neue Lochbandstück gelegt und so zusammengeklebt. Wenn das eingeschobene neue Lochbandstück lang genug ist, wird das andere Stück des Originalbandes darunter gelegt und zusammengeklebt. Als brauchbarer Klebestoff erwies sich nach einigen Versuchen der Fischleim.

5.4.3. Da das Lochband mit verschiedener Programmierung ziemlich vielmal (8 bis 10-mal) durch den Abtaster gehen muss, machen wir vom verbesserten (geklebten) Lochband immer eine Kopie, die entweder gar keine oder nur kleinere Verbesserungen hat. Dazu lassen wir das verbesserte Lochband durch einen Sender ("gépi adó") gehen, die Kopie wird durch die Klarschrift gleichzeitig oder nachträglich geprüft.

5.5. Programmierung des Automaten für die Textanalyse. Die Programmierung wird auf Grund des verbesserten bzw. durch den Sender und Fernschreiber kopierten (automatisch gelochten) Lochbandes vorbereitet.

5.5.1. Am oberen Teile des Automaten befindet sich die Programmtafel mit Einstellfeld und mit Überschrift der ersten und letzten Stelle der einzelnen Buchstaben bzw. (einen Laut bezeichnenden) Buchstabenverbindungen, Ziffern und Interpunktionszeichen. Es gibt an dieser Tafel für jeden Buchstaben, jede Ziffer, jedes Zeichen mehrere Löcher für die Stecker, um verschiedene Kombinationen (Verbindungen) stecken zu können.

5.5.2. Zur Forschung der Frequenz der einzelnen Buchstaben bzw. Lautzeichen verbindet man die Stelle des betreffenden Buchstaben (des Lautzeichens) mit einer der 49 Zähler (anfangs mit einer der 39 Zähler und als Ergänzung einer der nicht verwendeten 55 "letzten Glieder der Kombinationszähler" mit Überschrift UJ; vgl. 4.5.2. - 10 Zähler wurden in den Automaten nachträglich eingebaut, weil die 39 Zähler und die frei gebliebenen letzten Glieder der 55 Kombinationszähler oft nicht ausreichend waren. Die Frequenz der wichtigsten Zeichen (z.B. Punkt, Fragezeichen, Doppelpunkt) wird in der Regel an den freigebliebenen letzten Gliedern der Kombinationszähler gezählt.

5.5.3. Programmierung verschiedener Kombinationen. Am linken Drittel der Programmtafel findet man die Stecklöcher der Kombinationszähler. Sie bestehen aus 250 Einheiten: 55 Stecklöcher für die ersten Glieder der Kombinationen (bezeichnet E), 140 Stecklöcher für die Mittelglieder (K) - darunter 70 für den Anfang der Mittelglieder (KK), 70 für das Ende der Mittelglieder (KV) - und 55 Stecklöcher für das letzte (abschliessende) Glied (UJ) der Kombinationen.

5.5.3.1. Bei der Programmierung der Silbentypen dient immer die Position für Silbengrenze (/) zum ersten und letzten Glied. Zwischen diese Positionen (Anfang - / und Schluss - /) werden die Positionen für V (Vokal, Selbstlaut), C (K) (Konsonant, Mitlaut) und die verschiedenen Verbindungen

von diesen (z.B. CV, VC, CVC, CVCC usw.) als "Mitglieder" gesteckt.

5.5.3.1.1. Den Einheiten entsprechend, die am linken Drittel der Programmtafel miteinander verbunden sind, verbindet man im zweiten Drittel der Programmtafel die Positionen des ersten Gliedes (ED/) in bestimmtem Nacheinander mit den Positionen der Mit- und Selbstlaute (KD) und mit den Positionen des letzten Gliedes (UJ/). Der so geschlossene Stromkreis unterscheidet die einzelnen Silbentypen. Die entsprechenden Zähler des letzten Gliedes zählen die einzelnen Silbentypen.

5.5.3.1.2. Der Automat wurde ursprünglich so geplant, dass für Silbentypenzählen nur die ersten 18 Erstglieder zur Verfügung stand. Da es aber bei der Textanalyse nicht ausreichte, wurde die Verknüpfung so modifiziert, dass jedes Erstglied auch für Silbentypenzählen zur Verfügung steht.

5.5.3.2. Bei der Frequenz von anderen Buchstaben-, bzw. Lautverbindungen (wie bei verschiedenen Mitlautverbindungen) kann das Programm ähnlich vorbereitet werden, nur anstatt C und V verbindet man die einzelnen Buchstaben und Lautzeichen. Vom Zeichen der Silbengrenze (/) kann man natürlich absehen, wenn die Verbindung nicht binnen einer Silbe oder an einer Silbengrenze erforscht wird.

5.5.3.3. Für das Zählen von Länge der Wörter (nach Buchstaben bzw. Lautzeichen oder nach Silben), der Silben (nach Buchstaben bzw. Lautzeichen) und der Sätze (nach Wörtern) hat der Automat zwei spezielle Stromkreise: einen mit 21, einen mit 11 Zählern und mit je einem Summierer.

5.5.3.3.1. Mit dem Stromkreis mit 21 Zählern (AJ1 - AJ21) und einem Summierer (AT1) wird die Länge der Wörter und der Sätze, mit dem Stromkreis mit 11 Zählern (CJ1 - CJ11) und mit einem Summierer (AT2) aber die Länge der Silben gezählt.

5.5.3.3.2. Steckt man einen Stecker "Laut" ins Steckloch S1, einen anderen Stecker "Zwischenraum" ins Steckloch AT1, dann zählt AJ1 die Wörter, die aus einem Laut bestehen, AJ2 die Wörter aus 2 Lauten, .. AJ21 die Wörter aus 21 und mehr als 21 Lauten. AT1 zählt dann die Gesamtzahl der Wörter des Textes.

5.5.3.3.3. Beim Zählen der Länge von Wörtern den Silben nach verbindet man Silbengrenze / und S1. Bei der Länge von Sätzen nach Wörtern verknüpft man einerseits "Zwischenraum" und S1, andererseits "Punkt" (als Zeichen des Satzendes) und AT1.

5.5.3.3.4. Beim Zählen der Länge von Silben verknüpft man Buchstaben bzw. Lautzeichen mit S2, Silbengrenze / mit AT2. So zählt CJ1 die Silben, die aus einem Laut bestehen CJ2 die Silben aus 2 Lauten, ... usw. AT2 zählt dann die Gesamtzahl der Silben.

5.5.3.4. Zur Erforschung der Frequenz der Wörter ist im allgemeinen dieser Apparat nicht geeignet, da die Kombinationsmöglichkeiten relativ gering sind, da keine Ordnungsmöglichkeiten der Wörter und keine Schreib- oder Druckmöglichkeiten beim Automaten vorhanden sind. Nur einige der häufigsten kurzen Wörter kann man durch den Automaten zählen lassen, wenn man die einzelnen Buchstaben und Lautzeichen in entsprechendem Nacheinander zwischen zwei Zwischenräume programmiert (z.B. a, az, ám, egy, ez, igy, itt, ó, ott, ugy).

5.5.3.5. Die Programmierung der Wortartenwerte geschieht nach den einzel-
Zifferkombinationen der dreistelligen Zahlen. Hier kann man die erste

Ziffer zum ersten, die dritte Ziffer zum letzten Glied wählen, ohne Silbengrenzenzeichen und ohne Zwischenraum.

5.6. Das erste Abschreiben der Ziffern des Zählapparates ist notwendig um die so bekommenen Zahlen aus den späteren subtrahieren zu können.

5.7. Das "Lesen" des vorbereiteten Textes nach den einzelnen Programmen verwirklicht der Automat, wenn der rechts stehende Schlüssel des Abtasters in Grundstellung (d.h. Mittelstellung) gebracht wird.

5.7.1. Bleibt der Abtaster und der Automat an einer gewissen Stelle des Lochbandes stehen, so müssen Lochband und die betreffenden Stromleiter überprüft werden, um festzustellen, ob nicht etwa eine Leerstelle am Lochband stattfindet oder ob kein Kontaktfehler des Automaten das Stehenbleiben ausgelöst hat. Bei dieser Prüfung kann man auch Einzelschritte des Abtasters zu Hilfe nehmen. Das geschieht durch Rechtsstellung des rechten Schlüssels und durch einen einmaligen Druck nach rechts am linken Schlüssel und das sofortige Freilassen dieses zweiten Schlüssels.

5.8. Das zweite Abschreiben der Ziffern des Zählapparates nach Beendigung des "Lesens" zeigt das Plus des Textes gegenüber dem vorigen Zählerstand.

5.9. Die Subtraktion der Zahlen des früheren Zählerstandes aus den Zahlen des späteren Zählerstandes gibt die Zahl der Frequenz der programmierten sprachlichen Erscheinungen (des Buchstabens bzw. Lautzeichens, Interpunktionszeichens, des Silbentyps, des Wortartenwertes usw.).

5.10. Nach Tabellierung der Frequenzdaten des analysierten Textes erhält man die Verhältnisse und Proportionen zwischen den einzelnen sprachlichen Erscheinungen des Textes. Die Verhältnisse kann man entweder in absoluten Zahlen oder in Prozenten angeben. Die prozentuellen Angaben sind anschaulicher und mehr übersichtbar.

5.11. Wenn man verschiedene Texte analysiert hat, kann man diese verschiedenartig summieren: nach Autoren, nach literarischen Gattungen, nach Stilrichtungen, nach kronologischen Gesichtspunkten u.a.

5.12. Bei der sprachwissenschaftlichen Bewertung der Frequenzdaten muss berücksichtigt werden, dass die Frequenz der sprachlichen Erscheinungen eine sehr komplexe, sprachlich - gesellschaftlich - geschichtlich determinierte Gegebenheit ist, in der vielerlei Faktoren sich widerspiegeln und dass die Quelle dieser Forschungen d.h. der Text ein sprachliches Ergebnis individueller Tätigkeit ist, also neben dem "Sprachlichen", "Gesellschaftlichen" in hohem Grade auch "Individuelles" enthält. Damit hängt es zusammen, dass die verschiedenen Meinungen sprachstatistischer Untersuchungen in Einzelheiten oft so weit auseinandergehen. Es gibt aber auch überraschende Übereinstimmungen und die Erforschung der letzteren Tatsachen ist eben eine der wichtigsten Zielaussetzungen der sprachstatistischen Forschungen.

6. Für die weiteren Forschungen ist es also zweckmässig, die Prinzipien der Zusammenstellung eines Korpus festzulegen, sowohl hinsichtlich der Ausbreitung als auch der Gliederung, der Proportion der einzelnen Teile dieses Korpus nach Literaturgattungen, nach Verfasser und chronologischen Schichten der ausgewählten Texte. Erst nach einem mit grosser Umsicht zusammengestellten Korpus und seiner sprachstatistischen Analyse werden wir eine feste Basis zu den Teilarbeiten und weiteren Forschungen haben.

Diese Fragen und die sprachwissenschaftlichen Ergebnisse der Experimente wird Verfasser in einer späteren Arbeit eingehender erörtern.

A N M E R K U N G E N

[1] Über Plan und prinzipielle Fragen des Automaten: L. Kozma: Nyelvstatisztikai Automata (Ein sprachstatistischer Automat), Általános Nyelvészeti Tanulmányok (Studien über allgemeine Sprachwissenschaft), redigiert von L. Kalmár und Zs. Telegdi, II. Budapest, 1962-64. 133-7 p.

Über die sprachwissenschaftliche Grundlage des Automaten fand eine Besprechung zwischen L. Kozma und I. Fónagy statt; vor Beendigung des Automaten hat noch J. Kelemen einige Wünsche mitgeteilt, die teils auch eingebaut wurden.

Die technischen Arbeiten hat Béla Frajka, Dozent an der Technischen Universität und Ferenc Albert Projekttechniker der Fabrik für Fernmeldetechnik "Beloiannis" (Budapest) ausgeführt.

ОБЗОР ТАБЛИЦ ВЕНГЕРСКОГО СЛОВАРЯ, ПОЛУЧЕННЫХ НА ПЕРФОКАРТНЫХ МАШИНАХ

Ф. Пап

На страницах нашего журнала было уже сообщено о готовящейся обработке венгерского словаря на перфокартных машинах /см. Computational Linguistics III, стр. 205-211/. Ниже ознакомим читателей с таблицами, изготовленными за последние два года, а также с таблицами, намеченными до конца этого года. В приложение, в качестве иллюстрации, приведем также фотокопии отрывков некоторых таблиц. Предварительные результаты, полученные на основе этих таблиц, здесь не публикуются; они появляются и будут появляться систематически в соответствующих венгерских и заграничных органах. /Первая, вышедшая уже публикация, сообщающая наиболее общие данные относительно объема материала, распределения слов по длине, по количеству значения, по частям речи и т.д. Papp F.: Megszólaltak a gépek. - Magyar Nyelvőr, 1966. 2.

А. Напомним, что был обработан академический толковый словарь венгерского языка, итого - 58 323 слов. Рядом с каждым словом, путем применения простого цифрового кода, были кодированы следующие информации: сложное слово - несложное слово, омонимы, часть речи, количество значений, стилистические пометы, тип основы и окончания /для имен/, тип спряжения и сильное управление /для глаголов/, этимология, наличие суффикса - отсутствие суффикса. Кроме этого, механическим путем была установлена длина каждого слова в буквах. Таблицы, излагаемые ниже, построены на основе комбинации двух или больше кодированных информаций.

Система наших таблиц обладает некоторыми своеобразиями:

/а/ Изготовлены как полные таблицы с перечнем всех единиц, надлежавших упорядочению на машинах, так и сводные. Полные таблицы, пожалуй, не нуждаются в пояснениях. Если, положим, был дан приказ группировать слова по частям речи, то следуют сначала все глаголы словаря /глагол и имеет код 1 в столбце "часть речи"/, потом - все существительные /код:2/, прилагательные /код: 3/ и т.д. Из этой полной таблицы можно изготовить несколько разных сводных. Самая простая из них: привести только цифровые данные относительно того, сколько единиц было в разных группах по частям речи. Это - как бы "Содержание" на нескольких страницах шести объемистых томов, содержащих полные таблицы со словами, сгруппированными по частям речи. Можно даже написать рядом с каждой строкой страницу, где данная часть речи начинается. Уже и эта, самая простая сводная таблица дает в определенном смысле больше, чем полная: на сводной таблице в очень сжатом виде, вместе дано то, что разбросано по томам на полных списках. Эта сводная таблица поэтому в определенном смысле похожа на графики, ничем не отличающиеся от лежащих в их основе цифровых результатов, кроме сжатости, наглядности изложения результатов.

Однако мыслимы и более сложные сводные таблицы. Так, если названный только что приказ дополнялся приказом о том, чтобы, например, группировать слова в каждой части речи в порядке обратного алфавита /а приказ этого рода было необходимо издать, см. следующий пункт/, то можно требовать сводную таблицу, где бы внутри каждой части речи было указание на то, сколько единиц из данной части речи оканчивается на -А, на -В, на -С и т.д. Тогда перед итоговой цифрой, выражающей, положим, что в материале содержалось 30 574

существительных, содержатся цифры, указывающие на то, сколько из них оканчивалось на те или иные буквы, цифра 30 574 появляется как результат сложения частных результатов по буквам /окончаниям/. Далее. При желании узнать в известном смысле обратное по сравнению с этим, а именно: распределение частей речи по последним буквам /а не последних букв по частям речи/, в другой сводной таблице дается без труда и это: машиной слагаются результаты по последним буквам, частями этих результатов являются - количество слов той или иной части речи с данным окончанием.

В случае самых простых по структуре полных таблиц /см. табл. 1 и 2: словарь в алфавитном порядке и обратный словарь соответственно/ можно изготовить весьма интересные сводные. Так, можно потребовать, чтобы первые /последние/ две, три, четыре, пять букв были учтены постепенно, с последующими разбивками. Так, слова, начинающиеся с А, как общая сумма, разделяются на группы, в которых за А следует А, В, С...; группа АА соответственно разбивается на подгруппы, где после АА следует А, В, С...; группа ААА - на подгруппы АААА, АААВ, АААС и т.д. /Мы пошли только до первых трех букв, намерены пойти с правой стороны - т.е. с конца - до пяти букв, см. приложение 1./ Таким образом даются условные вероятности: если первая /последняя/ буква А, какова вероятность того, что последующая за ней буква будет А, В, С...; если первые две буквы - АА, АВ, АС и т.д., какова условная вероятность того, что третья буква будет А, В, С и т.д. /Точнее: даются, конечно, не сразу условные вероятности, но количества, на основе которых эти условные вероятности легко вычислимы./ Предполагается, что полученные таким образом результаты будут интересны со многих точек зрения. Укажем здесь только на то, что некоторые показатели Гринберга /например, показатели словосложения, префиксации, суффиксации/, пожалуй, более характерны, если их установить не на основе текстов, а на основе словаря. Ведь с точки зрения типологии существенным может оказаться не то, сколько раз названные явления повторяются в текстах, а то, в какой пропорции они представлены в словаре в целом.

Итак, первая особенность наших таблиц: наличие как полных /обозначаемых у нас буквой А после порядкового номера таблицы/, так и разных сводных таблиц /обозначаемых у нас буквой В после порядкового номера таблицы - если на основе одной и той же полной таблицы изготовлено несколько разных сводных, то они, естественно, нумеруются после В: 3.В-1, 3.В-2 и т.д./. Как видно из перечисленных примеров, сводные таблицы нередко не только суммируют содержание полных, но дают и новую информацию имплицитно заключенную в полных таблицах, но без соответствующих полных не явную. /А имплицитно все есть в обрабатываемом нами словаре, вся суть работы на машинах и заключается в том, чтобы сделать эту информацию явной./

/б/ Каждая классификация, проведенная нами, кончается тем, что дается указание, в каком порядке должны следовать друг за другом словарные единицы - в обычном алфавитном порядке /АВС/, или в порядке обратном /АТ, от латинского названия обратных словарей: a tergo 'сзади'/. Многие из классификаций проведены как в порядке АВС, так и в порядке АТ. Внутри строгого порядка АВС или АТ омонимы расположены по их порядковому номеру в столбце "омонимы".

В одном случае /7.2.В/ порядок АТ переплетается с упорядочением по грамматическим признакам: сначала берется последняя буква, потом - окончания, потом - предпоследняя буква и т.д., в порядке АТ. Эта, казалось бы сложная, классификация имеет очень прозрачный лингвистический смысл. А именно: рассматривается вопрос о том, при наличии каких предпоследних букв вы-

ступают те или иные окончания при одном и том же окончании /т.е. последней букве/ слова. Так, для общего случая: рассматриваются слова с последней буквой A. Оказывается, что у этих слов могут быть окончания либо типа X, либо типа Y, либо же типа Z. Окончания типа X выступают при условии, что перед конечным A имеются буквы B, C; окончания типа Y выступают при условии, что перед конечным A имеются буквы D , E и т.д. Для венгерских существительных подобная классификация чрезвычайно нужна, так как при агглютинирующем строе нет фиксированных окончаний для существительных /кроме случаев, когда существительное образовано с помощью суффикса/ в именительном падеже ед.ч.; в то же время при одной и той же последней букве могут выступать окончания разных типов. /Ср.: форма 3-го лица ед.ч.: comb-ja с окончанием -ja, láb-a - с окончанием -a при одной и той же последней букве -b./

Можно отметить, что, так как сортировка на машинах по ABC /AT/ довольно сложна /имеются в виду применяемые нами на этом этапе электромеханические машины/, специалисты предприятия Вычислительного центра Центрального статистического управления, где вся работа была проведена, приняли в этом вопросе особое решение. После того, как 58 323 карточки были готовы так, что на каждой карточке было написано слово дважды: с левыми полями и с правыми полями, между двумя написаниями с грамматическими информациями /см. Computational Linguistics III, стр. 210/, были изготовлены механически по две копии с каждой карточки, что на одной копии оказалось слово с левыми полями , на другой - только с правыми. После этого серия карточек со словами с правыми полями была классифицирована по AT, а серия карточек со словами с левыми полями - по ABC. Полученный на машинах порядок был нами проверен, ошибки в порядке исправлены. После этого карточки как в серии ABC, так и в серии AT были перенумерованы с 1 до 58 323. В дальнейшем каждая классификация была проведена либо на основе серии ABC, либо на основе серии AT так, что порядок ABC или AT был установлен на основе соответствующих порядковых номеров.

/в/ Некоторые классификации были проведены на всем материале, другие же - на тех или иных его частях, выделенных по особым критериям.

Важнейшей выделенной подгруппой оказалась подгруппа "Корневые слова". Под последними понимались слова несложные и не наделенные словообразующими суффиксами. Задача, казалась бы, простая - на самом деле лингвистическое решение ее оказалось весьма трудным. Есть несомненная группа корневых слов в указанном смысле - вроде láb 'нога', fa 'дерево', 'дрова', lát 'видит', szép 'красивый', ma 'сегодня', простые количественные числительные и т.д. /всего оказалось свыше 6 тыс. таких словарных единиц/. Есть, далее три группы слов, которые можно считать корневыми, но не так безусловно, как члены первой группы. Первую группу представляют собой слова с часто повторяющейся в словаре начальной частью /как: anti-, de-, ex-, pro- и т.п./. Если вторая часть таких слов оказывается регулярным элементом венгерского словаря, то такая единица считается сложным словом /напр.: exkirály 'экс-король','бывший король'/ и в группу корневых слов, естественно, не входит. Если же второго элемента нет в венгерском словаре /как: extempore, telegráf и т.п. - всего свыше 200 слов/, то слово считалось корневым /если, конечно, в него не оказалось суффикса/, но эти слова, в отличие от безусловных корневых слов, были выделены в особую группу.

В особые группы были выделены также слова с сомнительной конечной частью /fonológia, passziv /, а также слова, окаменевшие в венгерском словаре в форме какого-нибудь косвенного падежа - т.е. не наделенные суффикса-

ми, но и не без всякой грамматической морфемы в конце /közbe, midőn - всего свыше 2,3 тыс. и 1,4 тыс. единиц соответственно/. Итого в этой части содержится около 10 тыс. венгерских корневых слов, в каждой классификации отделены друг от друга названные группы корневых слов: безусловные корневые, с иностранным начальным элементом, с иностранным конечным элементом, с несуффиксом в конце. В дальнейшем, по ходу обработки материала, лингвист может работать с этими группами корневых слов в любой комбинации - может их всех считать вместе корневыми, или результаты той или иной группы причислить к результатам некорневых слов. В нумерации наших таблиц корневые слова имеют порядковый номер 31., после этого - номер соответствующей классификации и указание на характер А или В /см. выше/.

Кроме корневых слов были выделены глаголы /начало номера: 6./, имена /начало номера: 7. - внутри этого имена существительные, прилагательные, числительные, местоимения образуют особые подгруппы в каждой классификации; в особые группы входят также слова с характеристикой "существительное-прилагательное", "прилагательное-существительное" и т.п., т.е. слова, могущие принадлежать в разных контекстах к разным частям речи/. Были выделены также омонимы /10./, и некоторые другие группы слов. Некоторые выделения имеют смысл как для всего словаря, так и для корневых слов. Так, 31.6 - это корневые глаголы, 31.7 - корневые имена и т.д.

Б. Таблицы

/Цифрами 51, 52 и т.д. снабжены таблицы, еще не готовые в момент написания этого отчета/

На основе всего материала

1.А. Полный алфавитный список обработанного материала. - Механический алфавит несколько отличается от обычного алфавита, применяемого в венгерских лексикографических работах. /См. Magyar Nyelvőr, 1966, 2., ук. статья./

В. С материалом для вычисления условной вероятности до трех первых букв /см. приложение/.

2.А. Полный АТ список обработанного материала

В.-1. С разбивкой до пяти букв.

В.-2. С разбивкой до трех букв вместе взятых.

3.А. Длина, количество значений АВС

В.-1. Длина, количество значений.

В.-2. Количество значений, длина.

4.А. Часть речи, АВС

В. С разбивкой /внутри частей речи/ на начальные трехбуквенные сочетания.

5.А. Часть речи, АТ

В. как у 4.В /конечные сочетания букв/.

На основе глаголов только:

6.1.-А. Часть речи, тип спряжения, сильное управление АВ, АТ

Подразделение по частям речи нужно потому, что внутри глагола кодированы подгруппы по критерию "часть речи": безличные глаголы и т.п.

В. Как у 4.В

6.2.-А. Часть речи, спряжение, сильное управление, АВС

В. Как у 4.В. /В дальнейшем, если только особо не оговорено - части В. всюду - как у 4.В./

6.3.А. Часть речи, сильное управление, АТ

6.4.А. Часть речи, сильное управление, АВС

На основе имен только:

7.1.А. Часть речи, тип основы, окончания, АТ.

7.2.В. Часть речи, последняя буква, тип основы, окончания, предпоследние буквы. - Таблица А здесь не была заказана.

7.3.А. Часть речи, окончания, тип основы, АТ.

7.4.А. Часть речи, окончания, АТ.

На основе всего материала:

9.А. Стиль, АВС.

10.А. Стиль, АТ.

Только на основе омонимов:

10.1.А. АВС, омонимы.

10.1.В. Часть речи, АВС, омонимы.

На основе всего материала:

11.А. Часть речи, количество значений, АВС.

12.А. Сложное-несложное слово, суффикс, АВС.

13.А. Суффикс, сложное-несложное, АТ.

14.А. Сложное-несложное, этимология, АВС.

15.А. Сложное-несложное, этимология, АТ.

Только о простых словах /включая, в особой группе, и префикссальные образования/:

16.1.А. Сложное-несложное, стиль, происхождение, АВС.

16.2.А. Сложное-несложное, этимология, стиль, АТ. - По критерию "сложное-несложное" здесь можно было провести разбивку потому, что слова могут быть трех типов: простые, префиксальные и с сомнительным начальным элементом /иностранным/.

17.А. Часть речи, длина, количество значений, АВС.

Только на основе корневых слов:

31.1 - 31.11: Как на основе всего материала

31.12.А. Этимология, АВС.

31.13.А. Этимология, АТ.

31.14.А. Стиль, количество значений, АВС.

31.15.А. Часть речи, стиль, АВС.

31.16.А. Этимология, стиль, АВС.

31.17.А. Часть речи, длина, количество значений, АВС.

31.18.А. Длина, часть речи, количество значений АВС.

Кроме перечисленных здесь классификаций выполнен ряд упорядочений, с результатом - только сводными таблицами: 18.1.В - 18.12.В. /и, соответственно, на основе корневых слов также/.

51.А. Длина, АВС.

52.А. Длина, АТ. - Эти две классификации /без В, между прочим/, требуются не только потому, что, быть может, ради любителей кроссвордов можно было бы их издать /ср. аналогичную работу вышедшую во франции под редакцией Ж. Дюбуа/, но и потому, что они все равно необходимы в целях механического решения следующих классификаций.

53.А. Перечисление всех букв по позициям и сложение их по буквам и позициям. - Лингвистический смысл: сколько было всего букв и по какому распределению в обработанных словах /итак: сколько А, сколько В, сколько С и т.д./

54.А. Перечисление всех диграмм и их сложение.

55.А. Перечисление всех триграмм и их сложение. - Частный случай классификаций 54.А. и 55.А. - диграммы и триграммы в начале /соотв.: в конце / слов. Классификации 53-54. планируются и для корневых слов, с комбинацией критерия "этимология".

A	5
	5 *
ABA	2
ABA	2
ABB	6
ABC	2
ABE	1
ABL	24
ABN	2
ABO	4
ABR	9
ABS	13
	35 *
ACE	1
ACE	11
ACH	1
ACS	2
	15 *
AD	3
ADA	15
ADA	5
ADD	4
ADE	1
ADJ	5
ADM	5
ADO	13
ADS	55
ADR	2
ADT	3
ADU	2
ADV	3
	119 *
AER	2
	2 *
AFE	1
AFF	7
AFO	1
AFR	2
	11 *
AGA	6
AGA	1
AGG	23
AGI	5

Word	Comp.	Hom.	WCl.	Ms.	Style	Root	Endings Acc.	Endings Pl.	Endings Poss	Et.	Sf.	Length	2 a
TÖBBLET	1		2	2		1	04	04	04	9	9	7	
MUNKATÖBBLET	2		2	1	82	1	04	04	04	2	9	12	
ERTEKTÖBBLET	2		2	1	81	1	04	04	04	2	1	12	
JÖVEDELEMTÖBBLET	2		2	2	33	1	04	44	04	2	9	16	
ARTÖBBLET	2		2	1	82	1	04	04	04	1	9	9	
SEGEDLET	1		2	4		1	04	04	04	2	1	8	
RENDELET	1		2	3		1	04	04	04	2	1	8	
CSENDRENDELET	2		2	1	81	1	04	04	04	2	1	13	
SZÖKSEGRENDELET	2		2	1	82	1	04	04	04	2	3	15	
VÉGRENDELET	2		2	2		1	04	04	04	2	1	11	
FIÚKVÉGRENDELET	3		2	1	81	1	04	04	04	2	3	15	
MAGÁNVÉGRENDELET	3		2	1	81	1	04	04	04	2	9	16	
SAJTÚRENDELET	2		2	1		1	04	04	04	2	3	13	
KÖRRENDELET	2		2	1	33	1	04	04	04	2	1	11	
LAKÁSRENDELET	2		2	1	·	1	04	04	04	2	3	13	
SZABÁLYRENDELET	2		2	1	81	1	04	04	04	2	3	15	
KORMÁNYRENDELET	2		2	1	33	1	04	04	04	2	1	15	
LEHELET	1		2	4		1	04	04	04	2	1	7	
KELET	1	6	2	2		1	04	00	45	1	1	5	
KELET	1	7	2	1		1	04	00	05	1	1	5	
KELET	1	8	2	1		15	04	00	44	1	1	5	
KIKELET	6		2	1	49	1	04	00	04	2	1	7	
ESZAKKELET	2		2	2		1	04	00	44	2	1	10	
DÉLKELET	2		2	2		1	04	00	44	2	1	8	
NAPKELET	2		2	3	36	1	04	00	45	2	1	8	
LELET	1		2	4		1	04	04	04	1	1	5	
FELELET	1		2	2		1	04	04	04	1	1	7	
KINCSLELET	2		2	2	81	1	04	04	04	2	1	10	
LÁTLELET	2		2	1	33	1	04	04	04	2	1	8	
CSONTLELET	2		2	2	81	1	04	04	04	2	1	10	
TÉNYLELET	2		2	1	81	1	04	04	04	2	3	9	
ÉSZLELET	1		2	2		1	04	04	04	2	1	8	
EMELET	1		2	5		1	04	04	04	1	1	6	
FÉLEMELET	2		2	1		1	04	04	04	2	1	9	
BÖLCSELET	1		2	3		1	04	04	04	3	1	9	
JOGBÖLCSELET	2		2	1		1	04	44	04	2	1	12	
ÁLLAMBÖLCSELET	2		2	1	81	1	04	04	04	2	1	14	
VALLÁSBÖLCSELET	2		2	1	81	1	04	44	04	2	3	15	
TERMÉSZETBÖLCSELET	2		2	1	81	1	04	04	04	2	3	18	
NYELVBÖLCSELET	2		2	1	82	1	04	04	04	2	1	14	
VISELET	1		2	3		1	04	04	04	1	1	7	
MAGAVISELET	2		2	2		1	04	00	04	2	9	11	
HAJVISELET	2		2	1		1	04	04	04	2	1	10	
BÉRVISELET	2		3	1	60	1	00	07	00	2	3	11	
KÉPVISELET	2		2	4		1	04	04	04	2	1	10	
SZAKKÉPVISELET	3		2	2	48	1	04	04	04	2	1	14	
ERDEKKÉPVISELET	3		2	3		1	04	04	04	2	1	15	
KÜLKÉPVISELET	3		2	2		1	04	04	04	2	9	13	
NÉPKÉPVISELET	3		2	2	81	1	04	04	04	2	1	13	
VEZÉRKÉPVISELET	3		2	1	81	1	04	04	04	2	3	15	

12-B.	Comp.	Suff.	Quantity	
	1	0	6080	
	1	1	15321	
	1	2	3	
	1	3	8	
	1	8	1616	
	1	9	1425	
			24454	*
	2	0	8177	
	2	1	6719	
	2	2	2451	
	2	3	1575	
	2	7	1	
	2	8	328	
	2	9	1742	
			20993	*
	3	0	544	
	3	1	304	
	3	2	169	
	3	3	78	
	3	8	17	
	3	9	90	
			1202	*
	4	0	23	
	4	1	7	
	4	2	8	
	4	3	3	
	4	9	5	
			46	*
	6	0	1820	
	6	1	6807	
	6	2	1	
	6	3	3	
	6	9	1166	
			9797	*
	7	0	5	
	7	1	301	
	7	2	33	
	7	3	100	
	7	8	1	
	7	9	80	
			520	*
	8	1	10	
	8	2	1	
	8	3	7	
	8	9	1	
			19	*
	9	0	204	
	9	1	352	
	9	3	2	
	9	8	727	
	9	9	9	
			1294	*
			58324	*

Значение кодов:

Comp.	Suff.
1 = один корень	0 = суффикса нет
2 = два корня	1 = " в конце слова
и т.д.	2 = " внутри слохного слова
6 = один префикс + один корень	3 = суффикс и в конце, и внутри
7 = " + два корня	слохного слова
и т.д.	8 = псевдосуффикс /заимствованный/
9 = сомнительное словослохение	9 = сомнительный суффикс

ILLETVE	1		69	5		00	00	00	3	1	7			7225
												3		
VISZONTA6	1		69	4	12	00	00	00	1	9	9			7796
ILLET4LE6	1		69	5		00	00	00	2	9	9			8864
KÖVETKEZ4LE6	1		69	3		00	00	00	9	9	12			8868
AL16	1		69	7		00	00	00	1	9	4			10052
M26	1		69	5		00	00	00	1	9	3			10073
AM26	2		69	5		00	00	00	2	9	4			10074
												6		
26Y.	1		69	22		00	00	00	1		3			11173
												1		
CSAK	1		69	4S		00	00	00	9		4			13439
												1		
HOL	1		69	1S		00	00	00	1		3			21414
												1		
AM	1		69	14		00	00	00	9		2			23842
												1		
AZUTAN	2		69	4		00	00	00	2		6			25929
NEMKÜLÖNBEN	2		69	2		00	00	00	2	9	11			26025
MIKÉPPEN	1		69	5		00	00	00	1	9	8			26534
HATHISZEN	2		69	7		00	00	00	2	9	9			26652
HUN	1	2	69	1	45	00	00	00	1	9	3			27277
												5		
AKAR	1		69	9		00	00	00	1		4			34393
AMIKOR	2		69	6		00	00	00	2	9	6			36108
												2		
AMINT	2		69	S		00	00	00	2	9	5			52627
VALAMINT	2		69	9		00	00	00	2	9	8			52628
												2		
												22	*	
SZAMARA	1		S6	3		00	00	00	3	9	7			2811
												1		
HEGYIBE	1		SS	3		00	00	00	9	9	7			6351
RESZERE	1		S6	3		00	00	00	1	9	7			6953
												2		
HOSSZAT	1		S6	3		00	00	00	1	9	7			48683
ALATT	1		S6	13		00	00	00	1	9	5			53286
												2		
												582 5	*	

VALA	1	1	6	48	99	10	00	00	1		4	2003
VOLNA	1	1	2		99	10	00	00	1	9	5	2524
SINCS	1	1	16		99	20	00	00	2		5	5042
SOSINCS	2	1	1	79	99	20	00	00	2		7	5043
LABAD	1	1	2		1	10	00	00	1	1	5	5139
ZSIBBAD	1	1	3	24	1	30	00	00	2	1	7	5143
ELZSIBBAD	6	1	2		1	30	00	00	2	1	9	5144
PUFFAD	1	1	2	24	1	10	00	00	2	1	6	5156
MEGPUFFAD	6	1	1	51	1	10	00	00	2	1	9	5157
KIPUFFAD	6	1	1	51	1	10	00	00	2	1	8	5158
FELPUFFAD	6	1	2		1	10	00	00	2	1	9	5159
BEDAGAD	6	1	3		1	10	00	00	2	1	7	5161
MEGDAGAD	6	1	4		1	10	00	00	2	1	8	5162
KIDAGAD	6	1	3		1	10	00	00	2	1	7	5163
FELDAGAD	6	1	1		1	10	00	00	2	1	8	5164
BERAGAD	6	1	3		1	10	00	00	2	1	7	5171
LERAGAD	6	1	4		1	10	00	00	2	1	7	5172
ATRAGAD	6	1	2		1	10	00	00	2	1	7	5179
HIGGAD	1	1	3	51	1	30	00	00	9	1	6	5186
LEHIGGAD	6	1	2		1	30	00	00	2	1	8	5187
MEGHIGGAD	6	1	2	45	1	30	00	00	2	1	9	5188
VIGAD	1	1	2	78	7	30	00	00	9	1	5	5189
LEHORGAD	6	1	1	78	1	10	00	00	2	1	8	5198
LELOHAD	6	1	2		1	10	00	00	2	1	7	5208
ELKORHAD	6	1	2		1	10	00	00	2	1	8	5214
MEGPOSHAD	6	1	2		1	10	00	00	2	1	9	5217
KUSHAD	1	1	1		11	10	00	00	5	1	6	5218
LEKUSHAD	6	1	2		11	10	00	00	2	1	8	5219
ROTHAD	1	1	3		1	10	00	00	1	1	6	5220
LEROTHAD	6	1	1		1	10	00	00	2	1	8	5221
MEGROTHAD	6	1	1		1	10	00	00	2	1	9	5222
ELROTHAD	6	1	2		1	10	00	00	2	1	8	5224
RIAD	1	1	4		1	30	00	00	2	1	4	5226
FELRIAD	6	1	3		1	30	00	00	2	1	7	5230
SARJAD	1	1	2		1	10	00	00	9	1	6	5231
KIFAKAD	6	1	4		1	10	00	00	2	1	7	5239
ELFAKAD	6	1	2	45	1	10	00	00	2	1	7	5240
FELFAKAD	6	1	3		1	10	00	00	2	1	8	5241
ELAKAD	6	1	3		1	10	00	00	2	1	6	5243
FELAKAD	6	1	3		1	10	00	00	2	1	7	5244
FENNAKAD	6	1	6		1	10	00	00	2	9	8	5245
SZAKAD	1	1	12		11	10	00	00	1	1	6	5246
BESZAKAD	6	1	2		1	10	00	00	2	1	8	5249
FELBESZAKAD	6	1	1		1	10	00	00	2	9	11	5250
ÖSSZESZAKAD	6	1	3		1	10	00	00	2	1	11	5253
KETTESZAKAD	2	1	4		1	10	00	00	2	9	11	5254
MEGSZAKAD	6	1	2		1	10	00	00	2	1	9	5255
KISZAKAD	6	1	6		1	10	00	00	2	1	8	5256
ELSZAKAD	6	1	4		1	10	00	00	2	1	8	5257
FELSZAKAD	6	1	3		1	10	00	00	2	1	9	5258

ECS	2	1	04	00	04	2	
ICS	2	1	04	00	04	1	
						3	*
CCS	2	1	04	04	04	5	
ECS	2	1	04	04	04	6	
ECS	2	1	04	04	04	5	
ICS	2	1	04	04	04	7	
JCS	2	1	04	04	04	1	
NCS	2	1	04	04	04	16	
RCS	2	1	04	04	04	8	
						48	*
ECS	2	1	04	44	04	2	
NCS	2	1	04	44	04	1	
						3	*
CCS	2	1	05	05	04	2	
LCS	2	1	05	05	04	7	
NCS	2	1	05	05	04	1	
8CS	2	1	05	05	04	3	
4CS	2	1	05	05	04	11	
RCS	2	1	05	05	04	13	
7CS	2	1	05	05	04	1	
						38	*
CCS	2	1	05	05	44	1	
						1	*
RCS	2	1	05	55	04	3	
						3	*
ACS	2	1	23	00	00	1	
						1	*
ECS	2	1	24	24	14	1	
						1	*
						265	*
OCS	2	5	02	02	01	10	
						10	*
						10	*
8CS	2	99	20	00	04	1	
						1	*
8CS	2	99	55	55	04	1	
						1	*
						2	*
						277	*

SYNTHESE DES UNGARISCHEN HAUPTWORTES MIT EINER ELEKTRONISCHEN RECHENMASCHINE

Maria Stein

Der Algorithmus erstellt aus dem wörterbuchmässigen Stichwort auf Grund grammatischer Information aus der Analyse des Hauptwortes eine jegliche Form des Hauptworts. Die Aufgabe besteht somit aus der Verwandlung der grammatischen Informationen mit Hilfe (der unten angegebenen) regeln in eine Buchstabenreihenfolte, die an den Stamm des Hauptwortes angepasst die gewünschte Form des Hauptwortes ergibt.

Die Hauptbestandteile des Algorithmus sind die folgenden:

/A/ Ein Wörterbuch, das am Eingangsteil die abgekürzte Bezeichnung der einzelnen grammatischen Kategorien enthält, am Ausgangsteil jedoch die Buchstabenfolgen, die vom Algorithmus als Ergebnis geliefert werden können.

/B/ Ein Kontrollteil, der die grammatischen Informationen zwecks Anpassbarkeit auf ihre Richtigkeit hin kontrolliert.

/C/ Der sich auf den Stamm beziehendes Teil des Algorithmus.

/D/ Der die grammatischen Informationen umformende Teil des Algorithmus .

/E/ Vokalanpassung (Vokalharmonie).

/A/. Das Wörterbuch. In sprachwissenschaftlicher Abfassung gelangen wir vom Eingangsteil zum Ausgangsteil bei Verwendung der unten beschriebenen Regeln. Mit Hilfe dieser Regeln können die eingehenden grammatischen Informationen in eine Informationsreihenfolge umgeschrieben werden, die vermischt kleine und grosse Buchstaben enthält. Danach bei Anwendung der Regeln des Teiles /D/ auf die derart erhaltene Zeichenfolge gelangen wir zum Ausgangsteil des Wörterbuchs, der bloss kleine Buchstaben enthält.

I.	PL	K	11
II.	IPERS	M	21
	2PERS	D	22
	3PERS	É	23
III.	POSS	é	31
IV.	NOM	ø	41
	FOR	ként	42
	TEM	kor	43
	DAT	nEK	44
	CAU	ért	45
	SUB	rE	46
	DEL	rŐL	47
	INE	bEN	48
	ELA	bŐl	49
	ILL	bE	50
	ADE	nAl	51
	ABL	tŐl	52
	TER	ig	53
	ALL	hÖz	54

SUP	Ön	55
ACC	T	56
INS	WEL	57
FAC	WA	58

/B/ Man kontrolliert, ob die eingehende Informationsfolge reihenfolglich korrekt ist. Aus der obigen Beschreibung ist zu ersehen, dass die Informationen in vier Typengruppen geteilt wurden. Die Informationen verschiedener Typen dürfen jedoch nicht in einer beliebigen Reihenfolge aufeinander folgen. Steht eine Information des Typus k an der Stelle i der angebenen Informationsfolge, d.h. ihre ordnungszahl in der Informationsfolge ist i, so wird dies folgenderweise bezeichnet: s/k/ = i, wo k = I,II,III,IV, und i = 1,2,3 .... Die Reihenfolge der Informationsfolge ist richtig, wenn folgende Bedingungen erfüllt werden:

/1/ s/II/ < s/III/ < s/IV/

/2/ s/I/ < s/IV/

/3/ Informationen des gleichen Typus können nicht unmittelbar aufeinander folgen.

Wird eine der drei Bedingungen nicht erfüllt, so schreibt die Maschine "fehlerhaft" und sodann die fehlerhafte Informationsfolge aus.

Ist die Informationsfolge korrekt, so kommt ihre Erarbeitung an die Reihe.

/C/ Der sich auf den Stamm beziehende Teil des Algorithmus lässt den Stamm des Hauptworts unverändert, oder dehnt ihn (z.B. á statt a), oder ergänzt ihn durch ein j hinzu, oder dehnt und ergänzt ihn gleichzeitig. Zur Untersuchung wird der letzte Buchstabe des Stammes (u_1) beachtet. Der letzte Buchstabe kann viererlei sein:

/I/	a e
/II/	á o ó u u é ö ő ü u
/III/	i i
/IV/	b c d f g usw.

Bezeichnung der Zeichen von I und II, sowie der Zeichen Y, A, Á, E, É, Ö, Ő, Ü, O im Teil /D/: V (Vokal) Bezeichnung der Zeichen von III im Teil /D/: i Bezeichnung der Zeichen von IV im Teil /D/: C (Konsonant oder Consonant).

Der sich auf den Stamm beziehende Teil des Algorithmus ("Stamm" = die im Wörterbuch auffindbare Form):

/1/ ist u_1 = I und $Suff_1$ = 41 o. 42 o. 43, dann + /a o. e/ = u_1 ($Suff_1$: erstes Glied der Informationsfolge)

/2/ ist u_1 = V und $Suff_1$ = 23, dann + j (wo u_1 auch ein Ergebnis von /1/ sein kann)

/3/ ist u_1 = i und $Suff_1$ = 23 o. $Suff_{1,2}$ = /11/u./21 o. 22 o. 23/,dann +j

/4/ ist u_1 = C und $Suff_1$ = 23 o. $Suff_{1,2}$ = /11/u./21 o. 22 o. 23/,dann+/j/

/D/ ist der die grammatischen Informationen umwandelnde Teil des Algorithmus

/1/ K —→ I, wenn M o. D o. E auf K folgt (d.h. KM —→ IM, KD —→ ID, usw.)
K —→ i, wenn es nach é steht (d.h. éK —→ éi)
K —→ k, wenn es auf M folgt (d.h. Mk —→ Mk)
K —→ Ak in einem jeden anderen Fall

/2/ M —→ Yn, wenn auf M k folgt (d.h. Mk —→ Ynk)
M —→ Am in einem jeden anderen Fall

/3/. D→ötö, wenn auf D Ak folgt (DAk →Dötö)
D→Ad in einem jeden anderen Fall
/4/ I→Ei, wenn C vorangeht (CI→CEi)
I→i in einem jeden anderen Fall
/5/ T→ /A/t, wenn der Stamm vorangeht
T→Et in einem jeden anderen Fall
/6/ V_2→Ø, wenn V_1 vorangeht und wenn V_2 = i oder é (d.h. z.B. ea e, da V_1 = e, V_2 = a, welches vernichtet werden muss, öA →ö, usw.)
/7/ Á→á/é
/8/ Y→u/ü
/9/ A→O/Ü
/10/ E→a/e
/11/ ö→o/Ü
/12/ É→Y, wenn auf É Ak folgt (ÉAk→YAk)
É→á/é, wenn auf É die Information 31 o. 44 bis 58 folgt
É→a/e in einem jeden anderen Fall
/13/ ö→ó/ő
/14/ Ü→e/ö, wenn der Stamm vorangeht
Ü→e in einem jeden anderen Fall
/15/ ö→a/o
/16/ W→C, wenn C vorangeht (CW→CC)
W→v in einem jeden anderen Fall

/E/ Vokalanpassung (Vokalharmonie): Beseitigung der in den oben angeführten Schritten erhaltene Doppelergebnisse (u/ü, O/Ü, usw.):

/1/ Ist der letzte V = e, dann bei den Ergebnissen unter 7 bis 13 fällt der Buchstabe nach dem Bruchstrich weg, und zwar ü, e, é, ő.

/2/ Ist der letzte V = und i, so fällt der links vom Bruchstrich stehende Buchstabe weg (u, a, usw.).

/3/ Ist der letzte V = i, dann soll der im Stamm dem i vorangehende V untersucht werden:

3.1. Ist der dem i vorausgehende Vokal e oder i, dann bleibt die rechte Seite.

3.2. Ist der dem i vorausgehende Vokal nicht e oder i, dann bleibt die linke Seite.

3.3. Geht kein Vokal i voraus (d.h. das Wort ist einsilbig), dann bleibt die zweifache Schreibweise (u/ü, a/e, usw.) bestehen.

ANMERKUNG:

Die Angaben werden im Telexkode gelöchert. Da es am Telex keine gedehnten Vokale gibt, weder ein ö noch ein ü, so wird die gewohnte Transkription verwendet: ö = oe, é = ee, ü = ueue, usw. Linguistisch ist es gleichgültig, ob ein V = e, oder das V als zweites Element eine e enthält. Aus diesem Grunde können die Fragen im Teil /E/ so leicht gestellt werden, ob V = e oder V ≠ e.

Das i muss eigens untersucht werden, da das i nicht entscheidet, was für ein Suffix dem Stamm angehängt wird. Aus diesem Grunde falls das Wort einsilbig ist und ein i enthält, kann es nicht entschieden werden, welches Suffix benötigt wird. Ist aber vor dem i ein anderes i im Stamm, dann verhält es sich wie ein e.

Im Verlauf der automatischen Verwirklichung der Aufgabe ist die soeben beschriebene Gliederung übergeblieben. Bei der mechanisierten Durchführung des das Wesen der Aufgabe darstellenden Teiles /D/ erwies sich eine weitere Gliederung als vorteilhaft.

Bei der Umwandlung der Informationsfolge kommen erst die Informationen an die Reihe, die zum Types I gehören, oder in der Eingangsfolge der Information unmittelbar vor einer Information des Typus I stehen. Im folgenden Schritt werden die noch nicht bearbeiteten Informationen des Typus II aufgearbeitet, und schliesslich folgen die Informationen der Typen III und IV. Die Prüfung der Informationen des Typus IV verläuft folgenderweise. Es soll entschieden werden, ob die in Rede stehende Information zu einer Klasse gehört, die unabhängig von der Umgebung in eine gewisse Endungszeichenfolge umgeschrieben werden können. Bei einer verneinenden Antwort soll geprüft werden, ob die der Information vorausgehende, umgeformte Zeichenfolge auf einen Vokal oder einen Konsonanten ausgeht. Im letzteren Fall muss bei gewissen Informationen noch untersucht werden, auf welchen Konsonanten sie ausgeht, da auch dies die Umschreibung der in Frage stehenden Information beeinflusst. Z.B. ist die in Frage stehende Information INS, so kann diese auf kEl, mEl, dEl ... usw. umgeschrieben werden, je nachdem der vorausgehende Konsonant k, m oder d oder sonstiges ist.

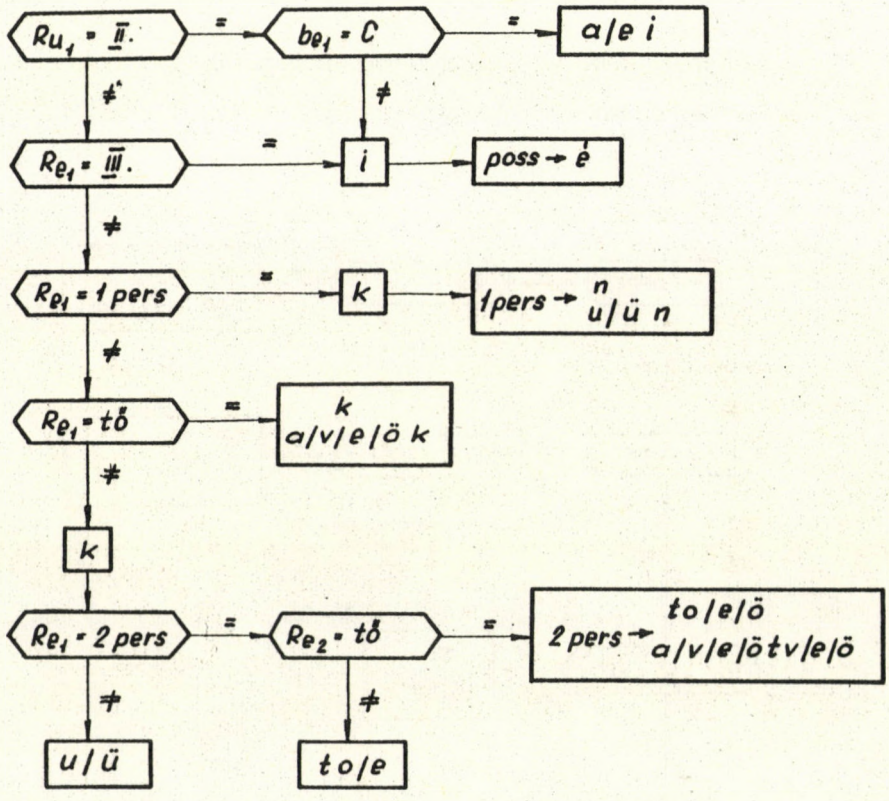
Die angeschlossenen Tabellen enthalten die Blockdiagramme des Teiles /D/ .

Eine Gliederung des Algorithmus ist zweckdienlich, da sie im weiteren Verlauf einen Zugang zum Program ermöglicht, sollen Erweiterungen oder Abänderungen vorgenommen werden.

Die Hauptwortsynthese kann auf verschiedenen Gebieten der sprachwissenschaftlichen Forschungen verwendet werden. So z.B. bei den automatischen Übersetzungsarbeiten, bei der mechanisierten Herstellung von Auszügen usw. Die Hauptwortsynthese ist ein Teil des Ganzen. Es wäre noch die Synthese des Zeitworts von Nutzen, wodurch zusammen mit der Hauptwortsynthese die volle ungarische morphologische Synthese zum grössten Teil verwirklicht werden könnte.

Der Algorithmus, wie eingangs angeführt, wurde von Ferenc Papp zusammengestellt. Ausführlich über die sprachlichen Beziehungen des Algorithmus s. bei Ferenc Papp: A magyar főnévragozás három modellje (Drei Modelle der Deklination der ungarischen Hauptwörter). (Magyar Nyelv, 1966. Nr. 2). Bei der automatischen Verwirklichung, die von Maria Stein durchgeführt wurde, wurde eine elektronische Rechenanlage URAL II der Rechenzentrale der Ungarischen Akademie der Wissenschaften verwendet, wo auch das ganze Synthesenthema ausgeführt wurde.

Erarbeitung der Informationen von Typ I. und der unmittelbar davor stenden Informationen



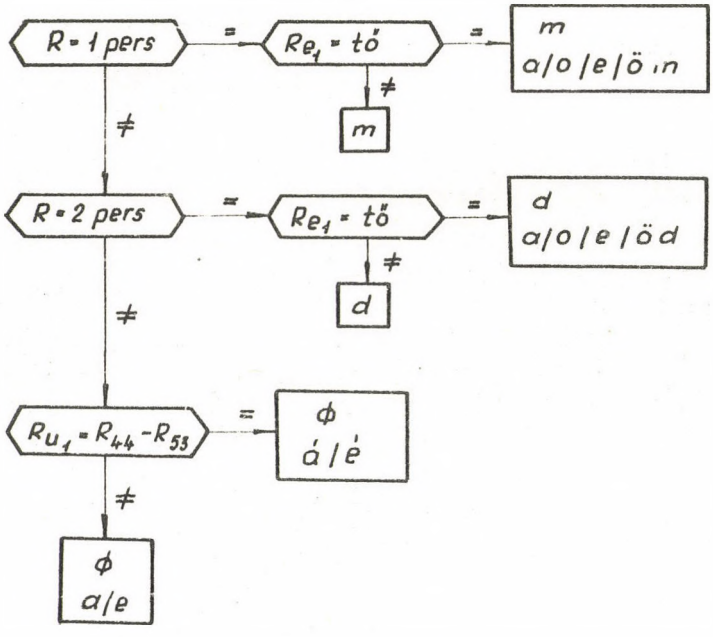
R = Suffix bzw. Information

b = Buchstabe

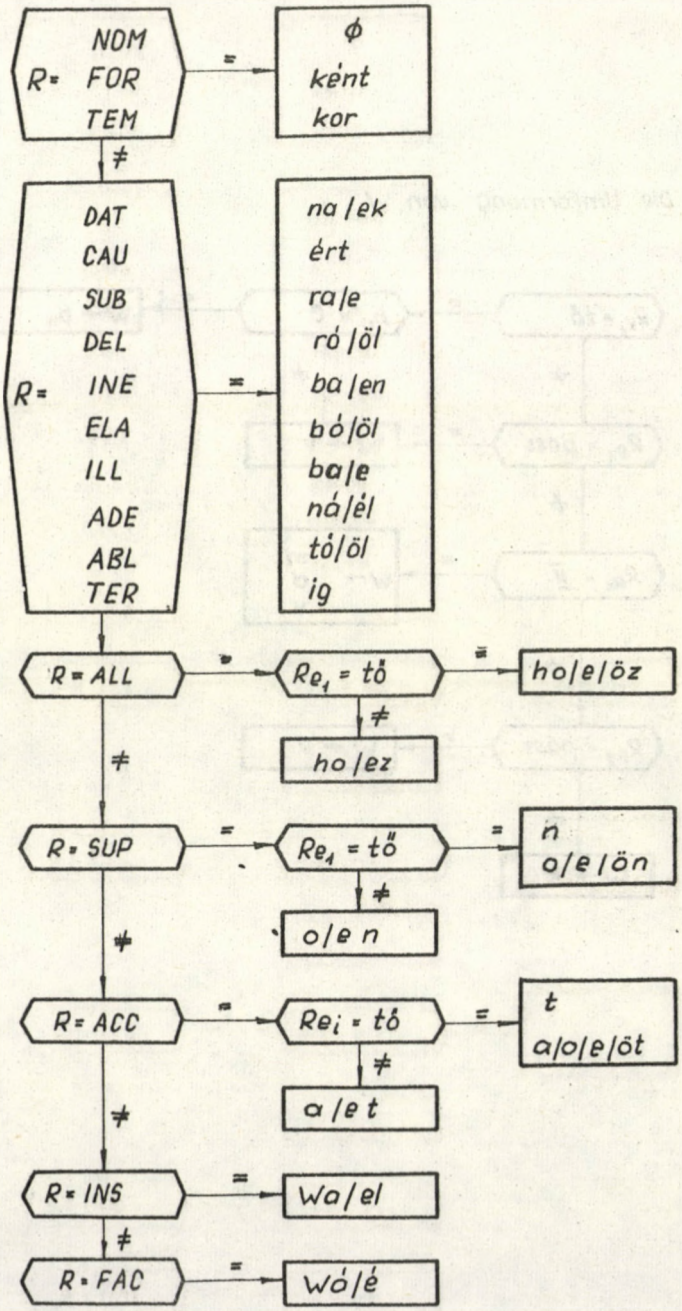
u_1 (Index) = die erste dahinter stehende

e_1 (Index) = die erste davor stehende

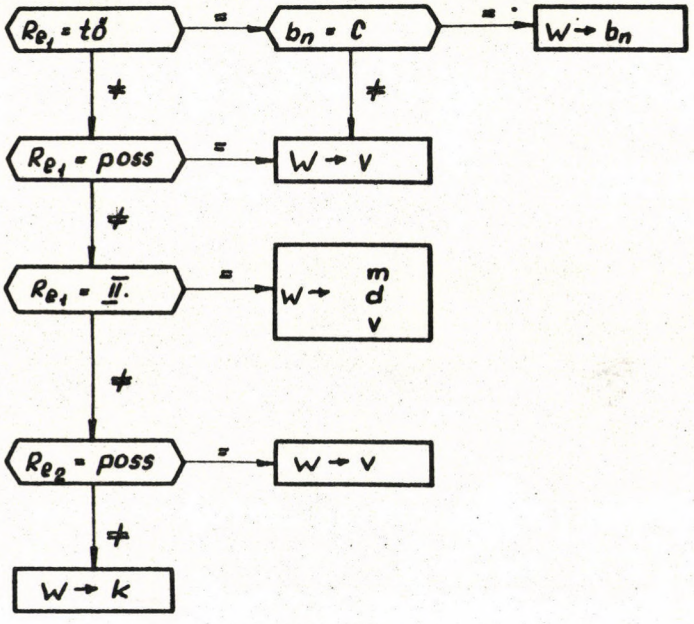
Die Erarbeitung der noch übrig gebliebenen Informationen von Typ II.



Die Umformung der Informationen von Typ IV.



Die Umformung von W .



REVIEWS

А. Людсканов:
ОСНОВИ НА ТЕОРИЯТА НА МАШИННАЯ ПРЕВОД С ОГЛЕД НА РУСКО-БЪЛГАРСКИЯ
МП. София. 224 р.

Bulgaria is one of those countries in which automatic translation could be taken up somewhat later than in some of the countries of the Western and socialist world. Although the technical conditions for automatic translation could be provided at a later date only, theoretical work had begun much earlier the proof of which is the present publication that was written - among several other articles - by Aleksandr Lyudskanov, the pioneer of automatic translation in Bulgaria. (For a brief summary of Bulgarian contribution to automatic translation see Computational Linguistics, Volume IV.)

The book sums up the results so far achieved in research work in the field of automatic translation and on reviewing an enormous amount of literary matter draws conclusions as to the notions and processes which may be utilized first of all in Bulgarian automatic translation. The work thus offers a thorough summary of the problems of automatic translation, and it is also suitable to acquaint those who know Bulgarian (many of those with a knowledge of Slavic languages) with the problem of automatic translation, above all with the linguistic aspects of automatic translation.

It is outside the scope of this review to go into the details of the contents of the work, partly because being linguists we do not feel competent to do so, partly because a detailed summary in Russian relieves us of this duty.

In the Introduction the author pays a great deal of attention to the peculiarities of automatic translation. The chapter dealing with automatic translation is one of the most interesting parts of the book, and at the same time it is pioneer work. The semiotic concept of translation which is developed by the author, permits the study of the theoretical problems of translation from an angle which appears to be the most appropriate from the point of view of linguistics.

The analytical section analyzes the problems of independent automatic translation, and discusses the general (26-47) and specific (48-187) linguistic problems of automatic translation. In the general part the problems of automatic translation are analyzed from the point of view of the theory of translation, in conjunction with what has been said in the introduction. Among others the problems of information required for automatic translation, "hundred per cent" translation, independent and dependent description, the intermediary language and the simplification of automatic translation (translation completely according to branches of science, preparation of the text) are discussed. The confrontation, evaluation and selection of the most appropriate of the various views is characteristic of the author's method of discussion. By this method he determines the objectives of Bulgarian automatic translation (first of all Russian-Bulgarian translation). It is hardly doubtful that this method is pertinent, naturally, whether the choice made without theoretical or experimental experience was correct, will be shown by practice only.

In the second part of the analytical section the problems of lexical,

grammatical analysis are studied. A separate chapter deals with the establishment of a uniform form of algorithms and the problems of translation according to meaning. We cannot undertake here to enumerate the problems and their solutions, those interested will find such an enumeration in the summary. The research and experimental methods of the author are characterized by a thoughtful, accurate analysis of his subject-matter. He defines on the basis of this analysis the practical procedure to be adopted. When dealing with a particular problem Lyudskanov surveys and utilizes the results of Bulgarian automatic translation. Primarily the chapter dealing with lexical analysis (among others the sections discussing the structure of a dictionary, the analysis of homonyms) deserves attention. Among the possible grammatical solutions the author often chooses the path that I. Melchuk took before him, with modifications he considers necessary. Chapter Four discusses the synthesis of the various algorithms and their uniform formulation, as a result of which "multiple meaning" may be eliminated by the same rules, however, with other constants on the various (lexical, morphological and syntactical) levels. The author discusses the problems of an analysis of meaning in a separate chapter, which is an indication of the fact that he is aware of the importance of this approach, although owing to the deficiency of the material at his disposal, the exposition of these problems is bound to be schematic.

The last part of the work bears the title "Synthetic Part". Here the author endeavours to define the organizational, methodological and linguistic principles of Bulgarian automatic translation, and to outline future research work, that is to promote directly or indirectly the success of Bulgarian automatic translation, first of all the success of Russian-Bulgarian translation, with special emphasis on the importance of structural and typological studies.

The work is completed by a copious bibliography and a detailed summary in Russian.

Reviewed by L. Dezső

Antal: CONTENT, MEANING AND UNDERSTANDING
Mouton & Co., The Hague, 1964. 63. p.

Antal's recent book takes up some of the ideas first formulated in his Questions of Meaning [1]. Here too, Antal accepts Morris's definition of meaning in its essence. The meaning of a sign is conceived as being identical to the rules of the use of a sign. Content reflects the relation between the linguistic sign and reality. Understanding refers to content and not to meaning.

Antal aims to segregate the spheres of linguistics and logic by pointing out that a sentence is not identical to a judgement, for though every judgement is a sentence, not all sentences are judgements.

In the following Antel investigates the connexion between content and understanding, and comes to the conclusion that the problem of content is not a subject-matter of linguistics.

The longest chapter of the book deals with the hypothetical character of meaning. According to Antal the meaning has to be postulated, otherwise we were unable to explain how a connexion is established between a linguistic sign and reality (the denotata of the sign). The hypothetical character of meaning appears to be confirmed by the fact that although meaning is part of the objective reality it cannot be grasped directly.

The last short chapter of the book discusses the ties between meaning and sentiment.

After this short summary of the book's content let us examine some of the arguable ideas of the author in some detail. It should be made clear however, that the reviewer does not want to be impartial, nor is this possible in this instance.

This statement should be understood in the following manner. Antal, in our opinion justly, turns against traditional linguistics and emphasizes the importance of the synchronic aspect in the theory of semantics as well. However, at the same time it is our feeling that Antal is not sufficiently well informed of the results and methods of formal linguistics. Or in other words, Antal in this work, like in Questions of Meaning, relies on inductive methods, although deductive methods would fit better into the present trends of research. Hence our prejudice has its origin in the fact that it is our intention to analyse Antal's work from the angle of the deductive theory of languages.

It is Antal's undisputable merit that he delimits semantics from logical semantics and psychologism. The value of the book lies more in its critical notes, than in its positive statements. In our opinion this is primarily due to methodological deficiencies.

Formal or structural semantics should be kept apart from general semantics. The aim of the general theory of semantics is to study problems which are on the borderline of linguistics and philipophy, or linguistics and psychology and to discuss general problems which because of their nature do not fit into a deductive theory, rather than to provide precise definitions for the various notions of semantics and still less to make efforts to describe the formal semantic aspects of language. Antal's book obviously deals with ge-

neral semantics. On the other hand, the formal or structural theory of semantics operates with accurately defined notions and strives for the formalization of some aspects (in our opinion, the most important ones) of the semantic properties of language. It can give rise to misunderstandings if no distinction is made between these two disciplines [2].

Antal believes that logic can hardly offer anything for the linguistic theory of semantics, in fact the subject matters of the two disciplines are wholly different. It is, however, apparent that many of the theories of mathematical linguistics are closely related to logic, first of all to symbolic logic. The categorial models of Y. Bar-Hillel and J. Lambek can be traced back to the semantic categories of the Polish logician K. Ajdukiewicz [3]. Furthermore, Bar-Hillel and R. Carnap outlined a theory of semantic information, which might be applicable to the study of natural languages as well [4]. To mention yet another logician, Curry's [5] influence on the development of mathematical linguistics is also far from being negligable. Naturally it is not question of direct application of the results of logic to the investigation of natural languages, still logical methods are often used to advantage. Antal's reasoning is quite correct if only a sort of general theory of semantics is meant by the theory of semantics. For deductive semantic models Antal's reasoning will hardly hold.

We depart from Antal's opinions also in the problem of grammatical correctness and semantic correctness. In our opinion this problem of correctness merely boils down to that of definition. As a matter of fact two alternative solutions offer themselves for clearing up the relation between grammatically and semantically correct sentences. One alternative is to build up a system which produces semantically correct sentences only (today the fact that semantic correctness presupposes grammatical correctness has received general approval). In this case semantic correctness is just a consequence of the recursive definition of "sentence". The other alternative is a system which produces grammatically correct sentences only, which means that in the generation of sentences no semantic information is used. That this should be feasible at all, the existence of two mutually exclusive sets of categories must be postulated. In this case semantic correctness must be defined separately. Moreover, in this latter instance only can an accurate definition of meaning be given [6]. Antal decides for the first of the two alternatives outlined above. Yet, as it has already been made clear, in this case no exact definition can be offered for meaning. Hence, though the question is by no means simple, it may be simplified, if it is narrowed down, i.e. an attempt is made to formulate it for the "formalizable part" of language only.

The next remark attaches directly to the foregoing. The notion of meaning as developed by Antal emphasizes the latent character of meaning. However, meaning may be made explicit, moreover, Morris's definition of meaning may take on an exact form. Naturally, the problem of meaning may hardly be considered settled by defining it in formal way. If we impose the requirement on our definition that it should fit into a more general notion of meaning, than we have succeeded in making an essential portion of meaning explicit.

One has to be extremely cautious when discussing the linguistic application of information theory. Although Antal tries to draw only a few conclusions from information-theory for the theory of semantics, yet it is much improbable that information theory has contributed much to linguistics. This holds all the more because information theory considers (or more precisely, is capable to handle the quantitative aspects of information only. What is usually called the information theoretic model of language, essentially does not offer anything new. All that may be said about it is that the old turns up in a new terminology. It is not our intention to dispute the practical results which have been achieved by defining the redundancy,entropy, etc. of language. We merely wanted to argue that it would not be justified, for the time being, to take over the theoretical conclusions of the theory.

As regards the morpheme, Antal is right when he emphasizes the need and importance of morpheme dictionaries. Mathematical linguistics has come to exactly the same conclusion, however, by another path. The definition of a morpheme is a totally different problem. It has only become evident during the last few years that linguistic units (phoneme, morpheme, word, sentence, to quote the most important ones only) cannot be defied by themselves, i.e. any definition will be inadequate unless it is provided within a well-defined system [7]. Hence it would be in vain to define a morpheme as the smallest unit of language conveying a meaning. This definition does not go far. The problems associated with this formulation of a morpheme are sufficiently known to make superfluous its repetition. On the other hand it would seem that a definition of linguistic notions can only be recursive (or what is equivalent to it, algorithmic), ilel a definition which allows for an enumeration or identification of all defined elements. However, this latter raises quite a few problems primarily because an excessive number of algorithms may be given for the identification of each a linguistic element. A satisfactory solution appears to be the one where the notions are defined within a uniform theory of language. The same applies to a number of other linguistic notions. In our opinion, a satisfactory definition has been found for both the phoneme and the morpheme within the generativ theory of language. The morpheme occupies a central position in Antal's reasoning and his opinion is strongly supported by the generative theory of language, although Antal's notion of a morpheme does not conform completely to that of the generative theory of language, and it is not even satisfactory in the light of what has been set forth above.

To sum up, Antal's work, notwithstanding its numerous merits, has in our opinion failed to enrich and consolidate the ideas which he refers to in his Questions of Meaning, although, as pointed out in the preface, the author himself has made this the main objective of his work. It remains our feeling that Antal's ideas lead into an impasse at many places, yet Antal has many ideas which may become useful for mathematical linguistics. To mention one, probably the most significant, Antal said on several occasions that meaning too is only a form, and that meaning can be studied through the form of the sentence. It would be very useful and instructive if Antal devoted time and energy to a more formal and positive study of meaning, and would not confine himself solely to the criticism of various semanticists. By this we wish to emphasize

we should like to hear a little more of Antal's own concept of meaning.

REFERENCES

[1] Antal: Questions of Meaning, Mouton & Co., The Hague, 1963. 95 p.

[2] See for more details the first two chapters of S. Ábrahám, F. Kiefer: A Theory of Structural Semantics, Mouton & Co. (forthcoming)

[3] On categorial grammar see: Általános Nyelvészeti Tanulmányok (Studies in General Linguistics), 3. 1965. 97-116 p.

[4] Y. Bar-Hillel, Y.R. Carnap: An Outline of a Theory of Semantic Information, Reading, Mass, 1964. 221-274 p.

[5] See e.g. H.B. Curry, R. Feys: Combinatory Logic, Amsterdam, 1958.

[6] Cf. F. Kiefer, S. Ábrahám: Some Questions of the Formalization Linguistics, Linguistics, 17. 1965. 11-20 p.

[7] See N. Chomsky: Categories and Relations in Syntactic Theory, in N. Chomsky: Aspects of the Theory of Syntax, Cambridge, Mass., MIT Press, 1965. X+251 p.

Reviewed by F. Kiefer

Р.М. Фрумкина: СТАТИСТИЧЕСКИЕ МЕТОДЫ ИЗУЧЕНИЯ ЛЕКСИКИ
Москва, 1964. 114 p.

"As a rule the operation of the linguistic units depends on so many factors, that for practical purposes it is impossible to take them all into account and to determine the trend of their interaction"... "It also stands to reason that an excessive accumulation of rules is unwelcome both for theoretical description and practical purposes". Setting out from these two statements Frumkina, in the appendix of her book, analyses a few problems of the justification of the establishment of statistical regularities, their conditions and sphere of validity.

She discusses sampling and the definition of the relative error in detail. (It is known that laws valid for large multitudes are established by sampling, however, the laws will hold only within certain limits. The size of the permissible error will in all cases depend on the problem under study.) Frumkina rightly points out the mistake so common among linguists namely that it is possible to determine the necessary size of the sample before the commencement of the analysis. This unfortunately is not the case. I feel that it would be useful to review this section of her book in greater detail.

The relative error may be determined from the formula

$$\delta = \frac{Z\varphi}{\sqrt{N \cdot p}}$$

(here δ denotes the relative error, N the number of elements of the sample, p the frequency of the phenomenon under study, $Z\varphi$ is the constant, which referred to the level $\varphi = 0.95$ equals 2. The level $\varphi = 0.95$ indicates that for hundred samples of elements N selected at random the permissible margin of error $P \pm \delta P$ can be exceeded in five instances only.) This formula contains two unknown terms at the outset of each test, i.e. p and N. Since it is exactly p that should be determined from the sample, it is obvious that N cannot be given beforehand.

Consequently a sample of N elements chosen at random should be made the starting point of the study. The sizes of the respective relative errors should be determined on the basis of the frequency of the specific phenomena under study, and the value of N should be calculated for that frequency whose relative error departs most from the value given by us. This gives us the required size of the sample. This process is demonstrated on an example. The strong point of Frumkina's book is the extremely lucid and yet rigorous investigation of the various problems.

Starting from the underlying principles summed up in the Appendix she evaluates the various dictionaries of frequency, and gives a summary of the methods for the correct compilation of such dictionaries and glossaries. Since in this sphere of problems G.K. Zipf's law is of fundamental importance, a chapter of Frumkina's book deals with the origin of this law and the attempts made to amend it. On the bases of the investigations made by J. Estoup (Gammes sténographiques, Paris, 1916), E. Condon (Statictics, of Vocabulary, Science 67 /1928/, p. 1733 et ssq.) and our own investigations Zipf has stated the

general validity of the law

$$p_r = k.r^{-\gamma}$$

(The Psycho-Biology of Language, Boston, 1935). In the law p_r denotes the frequency of a word referred to a given text, i.e.

$$p_r = \frac{\text{number of occurrences of the word in the text}}{\text{total number of words of the studied text}} = \frac{f_r}{N}$$

r denotes the serial number of the word in the glossary (the glossary is compiled so that the words are listed in the decreasing order of frequency, and within this, words of equal frequency are arranged in an alphabetic order and provided with a serial number from 1 to L), k and γ are constans (according to Zipf $k = 0.1$ and $\gamma = 1$). A critic of Zipf (M. Joos. Language, 12 1936) demonstrated that the "constants" were either those given by Zipf, and would accordingly apply only to texts where the number of the different words occurring in it was the same (L = 1200C), or they were not constants, both as the value of γ depended on the text. In a later work (Human Behavior and the Principle of Least Effort, Cambridge, Mass., 1949) Zipf essentially adhered to his earlier statement. Another dificiency of the law is its failure to reflect the internal and reciprocal ratios of the groups of the text put together of words of equal frequency. D. Mandelbrot tried to improve the accuracy of the law by introducing another constant $/p_r = k(r+\varphi)^{-\gamma}$ (Structure formelle des textes et communication, Word, 10, (1954) 1-28), however, in Frumkina's opinion this amendment doesn't bring about any fundamental changes. By taking all this into account she sums up the significance of Zipf's law as follows:

In the interval $50 < r < 1500$ the law in general reflects the true probability distribution of the words (the 1500 words of highest frequency represent about 80 per cent. of all words occurring in the texts), however, in the interval $1 < r < 50$ the law has to be amended. With this amendment Frumkina writes down the law as follows:

$$\sum_{r=1}^{B} kr^{-\gamma} = k\left(\text{const} + \sum_{r=n+1}^{B} r^{-\gamma}\right)$$

where $\text{const} = \sum_{r+1}^{n} p_r$ $(n \approx 50)$ may be accepted as constant for the language actually studied, and for $n < r < B$ γ is constant.

For the graph representing the law it is of little account whether in the dictionary of frequency each lexeme or lexicographical unit, or each differing word form is considered a "separate" word.

Since the compilation of the first dictionary of frequency (F. Kaeding : Häufigkeitswörterbuch der deutschen Sprache, Steglitz bei Berlin, 1898) near 300 such dictionaries and glossaries were compiled for sixteen languages.

According to Frumkina the functions of these dictionaries may be summed up as follows:

/a/ rationalization of the study of foreign languages (and the mother tongue);

/b/ improvement of the various code systems (e.g. shorthand);

/c/ study of the vocabulary of certain works of literature and writers.

Following the history of the origin of these dictionaries Frumkina briefly describes the particular stages of development grouping them by the functions indicated above. She discusses two dictionaries in detail, viz. Garcia Hoz. V., Vocabulario usual, comun y fundamental, Madrid, 1953 and H. Josselson, The Russian Word Count, Detroit, 1953, for the data only of these two may be used with scientific safety. As for the others these do not meet the conditions as regards sampling, nor are the values of error indicated in them. Consequently only a fraction of the words taken from these dictionaries reflect the actual conditions with accuracy.

At the compilation of dictionaries of frequency the following considerations have to be brought into harmony:

/1/ the dictionary should take up a substantial portion of the words;

/2/ the frequency of the rarest word included in the dictionary should be higher than an appropriately selected lower limit (let this be denoted by p_B);

/3/ for both theoretical and practical considerations the sample should not be excessive.

(The generally accepted relative error is $\delta = 0.3$).

On the basis of experience gathered in practice and in the analysis of the dictionaries it is sufficient for the dictionary to incorporate 70 to 80 per cent. of the words. (Above this level even a slight increase of the ratio entails an increase of the sample out of all proportions.)

When this ratio is given the value of p_B may be established, and with the knowledge of p_B and the given δ the number N of the elements of the sample may be determined conveniently from the formulas that Frumkina has described.

At the end of this chapter Frumkina briefly analyses the material of the dictionary recently published by Steinfeld:(E.A. Steinfeldt, Tchastotnyi slovar sobremennovo russkovo literaturnovo yazyka (2500 naibolee upotrebitelnikh clov), Tallin, 1963).

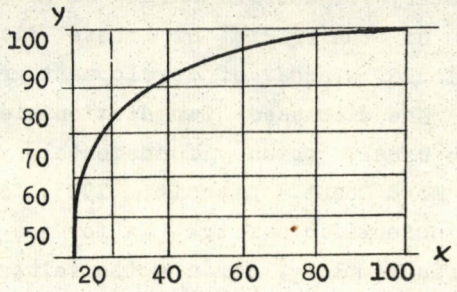
In the third and fourth chapters of the work the author deals with the analysis of the statistical structure of various texts.

The statistical structure of a text is known if to a random frequency of words the probability of occurrence (i.e. the probability of distribution) may be given.

G.U. Yule was the first who took an interest in the study of the statistical structure of texts (The Statistical Study of Literary Vocabulary, Cambridge, 1944). His primary purpose was to develop a suitable method for the establishment of the authorship of a work. To this end he looked for characteristics of the given text, yet, independent of its length. Although Frumkina doubts the value of these characteristics for a lexicon, still she suggests to try out the method on a Russian text of any length.

In the following she analyzes the stock of words of the Pushkin dictionary (see Maternaly k tchatotnomu slovoryu Pushkina, Moscow, 1963. 52.) (The stock of words of the Pushkin texts amount to 544 777 words, that of the dictionary (glossary) to 21 197 words). It is extremely instructive to observe the distribution of words by their frequency:

y: Volume of the text in percentages



x: Stock of words of the dictionary in percentages

More than one half of the words of the dictionary occur three times, or even less than that. (These should be subjected to a detailed qualitative analysis). The 200 most frequently occurring words account for 52 per cent. of the text.

This fact is significant particularly when it comes to a comparison of texts. (Garcia Hoz has quite correctly made it one of his principles that his dictionary should preferably contain words which are of similar occurence in the various types of texts). For sutdies of this type the method of computation of correlation is used in mathematical statistics. This method deals with the calculation of various coefficients.

As it is known the correlation coefficient is a number characteristic of the difference between the frequencies of given words in two different texts.If the reciprocal occurrences are plotted superimposed in a graph, the so-called correlation field, even those words may be ascertained which weaken the correlation. In many cases the correlation of degrees is the correct method to use for calculations. This serves for a comparison of objects of a different character. The author illustrates the application of both types of correlation on an abundance of examples.

The Appendix contains two lists compiled from the thousand most frequently occuring words in texts by Pushkin. The first list gives the frequency of these words, the second gives the distribution by types of texts of the most frequently occurring words of a few parts of speech.

Notwithstanding the large number of publications in this sphere of studies Frumkina's work will no doubt benefit the students of statistical linguistics.

Reviewed by S.J. Petőfi

J.C. Catford: A LINGUISTIC THEORY OF TRANSLATION
Oxford University Press, London, 1965.

Translation has been one of the most wide-spread means for transmitting intellectual works for several centuries. In the age of technical development its significance has grown to such an extent that today no research work can be done without knowing the scientific results achieved in foreign countries; that is why it is required to translate a considerable part of foreign scientific publications coming to light in a constantly increasing number. It is not by chance that almost immediately after the invention of electronic computers research work on the possibilities of translating by means of these machines of great working speed was started and has been carried on ever since.

The fundamental factor in machine translation is the formal analysis of language. The theoretical basis of this analysis is given by the grammatical systems of modern linguistics. But research workers had to realize very soon that there was no grammatical system by the aid of which any language could be described in a way desirable for machine translation. This recognition has resulted in publications demanding more exact description of language and efforts have been made to construct formal semantical systems which would make an exact description of language possible.

Machine Translation - by reasons of its task - cannot remain within the frame of a single language: being an operation of translation, its problems must be solved on the basis of similarities and differences between languages. What are these problems? There are a great many theoretical questions to be answered here: First of all: What is translation? Is there adequate translation at all? Are there and if so, what are the criteria on the basis of which the limits of possibilities in translation can be defined? What is the role of grammar and lexis in translation? Which unit of language should be considered as fundamental in translation? etc., etc.

The authors of various theories of translation have been answering these and similar questions in various ways according to their views about language. These theories of translation are generally incomplete: they do not cover every problem of translation and are mostly intuitive, insofar as they do not take a given linguistic theory for their basis.

One of the greatest merits of the theory of translation advanced by J.C. Catford is that it has been built on a definite, elaborated linguistic theory. The author examines every part of it from the viewpoint of translation and proves his statements with the theses of the linguistic theory taken as his basis. His work can serve as an example for other authors, because it shows how an exact translation theory can and must be built on a given linguistic theory.

In the compass of a short paper like the present one it is impossible to review every detail of the theory; it seems more profitable to concentrate on those parts which contain statements of a theoretical significance, on the one hand, and instructive for Machine Translation, on the other.

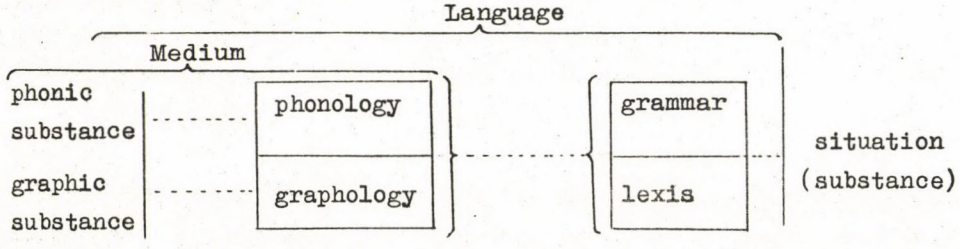
J.C. Catford is one of the leading members of the Edinburgh Linguistic

School and is a prominent linguist dealing with both general and applied linguistics. His theory of translation is based on the linguistic theory developed at the University of Edinburgh, in particular by M.A.K. Halliday [1] but takes into account the results achieved by J.R. Firth to a large extent. This linguistic theory can be summarized briefly as follows:

I. General Linguistic Theory

There are two extralinguistic levels in language behaviour: 1) phonic and graphic substance used as "means" by the performer and 2) situation substance to which communication is related. Language is nothing else than "the organization or patterning which language-behaviour implicitly imposes on these two kinds of substance - language is form, not substance" [2]. The internal levels of language are the two levels of medium form: phonology and graphology, and the two formal ones: grammar and lexis. These levels are the results of abstraction.

Language behaviour can be summarized as follows [3]:



The relationship between grammar/lexis and situation is the contextual meaning of the text [4]. Likewise the relationship between phonology and phonetic substance is phonetic meaning and that between graphology and graphic substance is graphetic meaning.

The fundamental categories of linguistic theory are: unit, structure, class and system.

1. Unit is a stretch of language activity which is the carrier of recurrent meaningful patterns. The units of grammar or of phonology operate in hierarchies: more inclusive units are made up of less inclusive ones. Units of English grammar form the following "ranks of hierarchy": Sentence, Clause, Group, Word, Morpheme. Statements about "ranks" are of great importance in the theory of translation.

2. A structure is an arrangement of some given elements. Thus, the elements of structure of the English clause are: P (predicator), S (subject), C (complement), A (adjunct).

3. A class is given by grouping the members of a unit according to the way in which they operate in the structure of the unit next above in the rank scale. E.g.: The class of Verbal Groups operate as P in clause structure; the class of Nominal Groups operate as exponents of S or C in clause structure, etc.

4. A system is a finite set of alternants, among which a choice must be made for a given role. An example in grammar might be the number system of nouns.

5. The above categories are not applicable to lexis: it is not a closed

system. The formal discussion of lexis is made in terms of collocation and lexical sets. E.g.: "sheep" collocates frequently with such lexical items as "field, flock, shear", etc., but its collocation can be "roast" as well. The different collocational ranges show that we have different lexical items. A lexical set is a group of lexical items having similar collocational ranges. If two lexical items are of the same medium exponent (graphological and phonetical), it is the lexical set that enables us to decide which of them is given in a certain case. E.g. the graphological form "bank" in English enters into two distinct collocational ranges, and so into two distinct lexical sets, so we can say that it denotes two lexical items of the same form (in Hungarian "bank", "part").

II. Translation Equivalence and Formal Correspondence

Translation is "the replacement of textual material in one language (SL) by equivalent textual material in another language (TL) [5].

1. When translating, the task is to find a TL form (text or portion of text) "wich is observed to be equivalent of a given SL form (text or portion of text)" [6]. A formal method for finding a translation equivalent is commutation. It means that we change a given portion of SL and observe the consequence of this change in TL. A textual translation equivalent is "that portion of a TL text which is changed when and only when a given portion of the SL text is changed" [7]. E.g.: If the English sentence: "My son is six" is translated into French, we get: "Mon fils a six ans". If the sentence is changed this way: "Your daughter is six", the French will be: "Votre fille a six ans". The changed part of TL (mon fils - votre fille) is considered as the textual translation equivalent of the portion changed in SL (my son - your daughter).

Sometimes there is a difference between the SL and TL equivalents in rank and structure. In such cases commutation must be applied for the whole clause (sentence). E.g.: Be given the following English SL text: "The woman came out of the house" and its Russian TL equivalent: "Женщина вышла из дому". Suppose we want to discover the Russian equivalent of the English definite article "the" in the group "the woman" of the given text. Commutation might give the following result:

SL text 1. The woman came out of the house.

TL text 1. Женщина вышла из дому.

SL text 2. A woman came out of the house.

TL text 2. Из дому вышла женщина.

So in this particular case the change of the definite article into the indefinite is correlated with a change in the sequence of elements in the structure of the Russian clause. Thus in a Nominal Group taking the place of Subject in a Clause Structure the result of the commutation is:

Eng. the in (N) at (S) = Rus. (SPA).

Eng. a in (N) at (S) = Rus. (SPA).

The method of commutation is applicable also in cases when the given SL item has no equivalent in the TL. In such cases we say that the equivalent is "nil", reserving the term "zero" for elements working in the system of a language but lacking in a given text. E.g.

SL Eng. My father was a doctor.

TL Fr. Mon père était docteur.

TL Rus. Отец у меня был доктор.

In this example the equivalent of the English indefinite article "a" is "zero" in French (there is indefinite article in the French language as well), but in Russian, which has no system of articles, the equivalent is "nil". In this example equivalence can be established only at a higher rank, namely at the rank of Group: the equivalent of the English nominal group "a doctor" is the Russian nominal group (of one member) "доктор". In general,nil equivalence implies that equivalence can be established only at a higher rank.

If the TL equivalent of a given SL item is the same at each occurence, we say that the probability of the occurence of X_1 in TL is 1 in any case when X occurs in SL; and it is 0 when X_1 never occurs in TL when X occurs in SL. Probability may have any value between 1 and 0. The probability values of textual translation equivalences can be generalized in the form of translation rules. For the purpose of Machine Translation these rules may be operational instructions (algorithms). It is quite natural that "correct" results can be got only when the probability value is 1 (or at least very near to it).

2. Formal correspondence means that a category in TL occupies as nearly as possible the "same" place in TL as the given SL category in SL. E.g. there is formal correspondence between the word-classes preposition in English and French because the morpheme called "preposition" is to be found in both languages in nominal groups functioning as adjectives in nominal group structures (the door of the house - la porte de la maison) and as adverbs in clause structures. But this statement makes it necessary to justify the formal correspondence between the units of higher rank (nominal group structure, clause structure, etc.); and this can be done on the basis of textual equivalence. The degree of divergence between textual equivalence and formal correspondence may be used as a measure of typological difference between languages.

III. Meaning

It is generally agreed that meaning has great importance in translation. Consequently it is necessary for a theory of translation to draw upon a theory of meaning because without such a theory a lot of important aspects of the translation process cannot be discussed.

The author derives his theory of meaning largely from the views of J.R. Firth. According to this theory meaning is a property of a language. In respect of translation it means that a given SL text has its SL meaning and a TL text its TL meaning, i.e. a Russian text has a Russian meaning (as well as Russian phonology/graphology, grammar and lexis), and an equivalent English text has an English meaning. This is a clear consequence of the definition of meaning given by Firth, according to which meaning is a "total network of relations entered into by any linguistic form - text, item-in-text, structure, element of structure, class, term in system - or whatever it may be" [8]. The relations entered into by units of grammar and lexis are of two kinds: formal and contextual relations.

1. Formal relations are ones between a given formal item and others in

the same language. In grammar this may be the relation between units of different rank, the relation between a class and an element of structure at a higher rank, etc. In lexis there are formal relations between one lexical item and others in the same lexical set, and formal co-textual relations between lexical items in texts.

2. Contextual relations are the relations of grammatical or lexical items to linguistically relevant features in the situation in which the items operate. The situational features which are contextually relevant to a given grammatical or lexical item are discovered - as translation equivalents - by commutation. If we change a feature of situation, a certain change will occur in the text; if we change an item in the text, a certain change will occur in the situation. Thus a range of situational features relevant to a given linguistic form can be established, and this range constitute the contextual meaning of that form.

3. It has already been mentioned that formal correspondence between languages can only be a rough approximation; consequently the formal meanings of SL items and TL items can rarely be the same. A dual item of a language having a 3-term number-system (singular, dual, plural) may be the translation equivalent of a plural item of another language having a 2-term number-system (singular, plural), but it cannot have the same formal meaning [9].

The same can be said about contextual meanings: the group of the relevant situational features of an item varies from one language to another. One of the examples by the author: A girl walks in and says: "I've arrived." The situation in which the text occurs - just as any other situation - is indefinitely complex, it has innumerable features (date, place, the age, name of the girl, the colour of her eyes, hair, her clothes, the number and nature of her audience, etc., etc.). However, only very few of these features are linguistically relevant, that is to say, built into the contextual meaning of the text and its parts. These are the following:

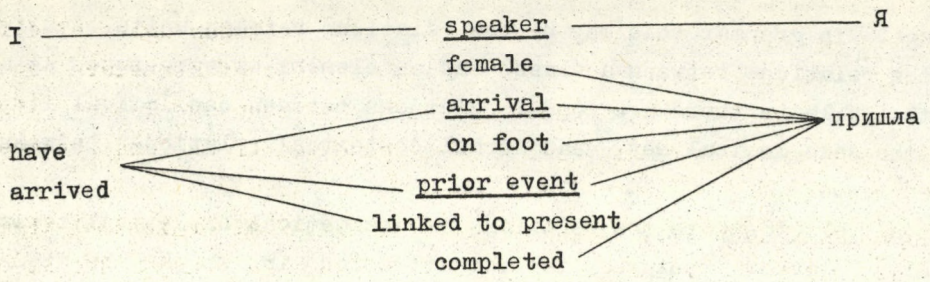
a) the speaker is the performer of the action (I),
b) the action is arrival (arrive),
c) a prior event which is
d) linked to a current situation (perfect form),
e) the current situation is in the present (Present Perfect), etc.

The sentence translated into Russian is: "Я пришла"

The linguistically relevant features now include:

a) the speaker is the performer of the action /Я/
b) the speaker is female /пришла/,
c) an arrival /прийти/,
d) on foot /прийти/,
e) a prior event (past),
f) completed (perfective verb-form) . . . etc.

Though the Russian text is a perfectly good translation equivalent of the English text, it does not "mean the same", because it selects - as linguistically equivalent - a different set of features in the situation. The difference can be illustrated in this way:



Only the situational features underlined in the list are contextually relevant to both the SL and the TL text. There are some among the situational features which are relevant only to the TL text (in the given example: female, on foot, etc.). These elements can be determined only by means of an analysis of a larger part of the SL text, i.e. on the basis of a larger contest [10].

IV. The Limits of Translatability

The above (III.) and similar examples show that definitions according to which translation is replacement of a SL item (text) by a corresponding TL item (text) of the same meaning, are linguistically unacceptable. A statement like this would restrict the possibilities of translation to the cases where the SL and TL items are linguistically equivalent, i.e. where all the linguistically relevant features of a given situation are common to both the SL and TL items (texts). But this is very rarely the case. Similarly rare are the cases where there is no linguistically relevant situational feature common in the two items. In most of the cases some of the linguistically relevant situational features are common, some are different. As we have seen, a given TL item must be considered as a perfectly good translation equivalent of a SL item even if it has only some linguistically relevant situational features common with the SL item. The number of the common elements, naturally, can be very different. That is why we should say that SL items or texts are more or less translatable rather than that they are absolutely translatable or untranslatable.

A given SL item (text) must be considered untranslatable when it has no linguistically or functionally relevant situational feature common with any TL item (text). Untranslatability has two causes: linguistic and cultural.

1. Linguistic untranslatability occurs when there are formal features in the SL to express the functionally relevant features of a given situation but in the TL the corresponding formal features are lacking. It occurs typically in SL puns, where an ambiguity peculiar to the SL is a functionally relevant feature. Ambiguities causing untranslatability arise from two sources: homonimy and polysemy. Because of the formal differences between languages ambiguities of these types are very frequent and lots of examples might be adduced; in most of the cases, however, the context of the ambiguous item contains those functionally relevant features of the situation, by means of which the translation equivalent can be determined.

2. A quite different problem arises when a situational feature, functionally relevant for the SL text, is absent from the culture of which the TL is a part. E.g.: There is no translation equivalent of the Finnish lexical item "sauna" in English because the institution is unknown on the territory where

English is spoken. There might be texts, in which "bath" or "bathroom" are acceptable translation equivalents, but these do not contain the functionally relevant features of the situation expressed in the SL item. In such cases most of the translators adopt the solution of transferring the SL item into the TL text unchanged and explaining it in a footnote.

Statistically "cultural untranslatability" occurs much more rarely than linguistic. But culture is a factor outside language and so linguistic problems connected with it - at least at the present stage of our knowledge - cannot be solved.

REFERENCES

[1] M.A.K. Halliday: Categories of the Theory of Grammar, Word, 17. 3. 241-92. and M.A.K. Halliday, A. McIntosh and P.D. Strevens:The Linguistic Sciences and Language Teaching, Longmans, 1964.

[2] p. 3.

[3] The fact that the author considers phonology and graphology as independent levels (Halliday calls them "interlevels" connecting phonic and graphic substances with the "formal levels" grammar and lexis) has very important theoretical consequences: this consideration makes it possible for him to work out the rules of translation restricted to the various levels ("phonological", "graphological", "grammatical" and "lexical" translation). As "restricted translation" is of less importance from the practical point of view, we are not going to touch upon its rules here.

[4] The word "meaning" as used in everyday life, is practically equivalent to "contextual meaning".

[5] p. 20.; SL = source language, TL = target language.

[6] p. 27.

[7] p. 28.

[8] p. 35.

[9] They give quite different information about the situation.

[10] There are two tendencies in the attempts for constructing formal semantics. One of them takes the "thesaurus system" as its basis, the other takes into consideration the features of reality (situation) connected with a word. Catford in his theory of translation uses both: he makes his statements about lexical sets on the basis of the "thesaurus system", while his rules of translational equivalence are based on the linguistically relevant features of the situation.

It will be worth mentioning here the Semantic Theory by Katz and Fodor. They, too, use "elements of meaning" (compare: situational features) and contextual investigations for determining the "meaning" of a word, although setting out from other starting-points. (J.J. Katz, J.A. Fodor: The Structure of a Semantic Theory, Language, 39. 1963. 170-210.)

Reviewed by Gy. Sipőczy

AUTOMATISCHE SPRACHÜBERSETZUNG
ENGLISCH-DEUTSCH

Die Arbeitsstelle für mathematische und angewandte Linguistik und automatische Übersetzung der Deutschen Akademie der Wissenschaften zu Berlin

Mitarbeiter: E. Agricola, J. Kunze, S. Nündel,
J. Stadelmann, I. Starke, F. Siegmund-Schultze. Berlin, 1965. 162 p.

Im Institut für mathematische und angewandte Linguistik der Deutschen Akademie der Wissenschaften wurde im April 1962 eine Serie von Experimenten auf dem Gebiete der automatischen Sprachübersetzung auf der elektronischen Rechenmaschine URAL I in Angriff genommen.

Die Ergebnisse der Experimente sind zum Teil konkrete Erfolge, welche mit der praktischen Verwirklichung der Übersetzung im Zusammenhang stehen. Die Ergebnisse sind zum grossen Teil bereits praktisch verwendbar, so z.B. die syntaktische und morphologische Synthese der deutschen Sprache. Ein anderer Teil der Ergebnisse sind grundsätzlicher Natur und beziehen sich auf die Möglichkeiten und Bedürfnisse der weiteren Arbeit.

Im folgenden werden die Schlussfolgerungen und Ergebnisse der Experimentalserie zusammengefasst:

/1/ Die Vorstellungen und Methoden der Arbeitsstelle sind durchführbar, die aufgekommenen Probleme sind lösbar.

/2/ Erkenntnisse über die im Übersetzungsvergang auftauchenden Operationen und deren zweckmässigste Reihenfolge.

/3/ Feststellung der Typen der in der Praxis auftauchenden Schwierigkeiten (lösbar; lösbar, jedoch noch nicht ausgearbeitet, usw.).

/4/ Angaben zur Planung und über Umfang der zu lösenden Aufgaben.

/5/ Vorschläge zur Beseitigung der Begrenzung der gegenwärtigen technischen Gegebenheiten.

/6/ Festsetzung der Grenzen der automatischen Übersetzung bei Verwendung der bisherigen formal-linguistischen Methoden.

Die Hauptabschnitte der Experimentalarbeit sind die folgenden:

/1/ Das englische Analysenwörterbuch
/2/ Bestimmung der Wortart
/3/ Die syntaktische Analyse
/4/ Umwandlung der Konstruktionen und Idiome
/5/ Die syntaktische Synthese
/6/ Das deutsche Synthesenwörterbuch
/7/ Die morphologische Synthese

Von diesen sind die wichtigsten und zugleich lehrreichsten die syntaktische Analyse und die syntaktische Synthese.

Im Verlauf der Analyse wurde bei Beachtung der technischen und praktischen Gegebenheiten eine Abhängigkeitsgrammatik verwendet. Er wurden die im Satz bestehenden mittelbaren und unmittelbaren Abhängigkeitsverhältnisse festgesetzt. Zur Formalisierung der syntaktischen Zusammenhänge wurde die Baumzeichnung verwendet, und zwar nach den Vorstellungen von Tesnière, Melčuk und Andreev.

Im Verlauf der syntaktischen Analyse wurde der Satz, d.h. eine Folge von Wörtern, in eine Symbolenfolge umgewandelt und im weiteren Verlauf wurde dann statt konkreter Wörter, mit dieser Symbolenfolge gearbeitet. Durch Anwendung gewisser Regeln wurde diese Symbolenfolge in eine neue Zeichenfolge umgewandelt. Dies wird derart vorgenommen, dass zu einer, je einem Satz entsprechenden, Symbolenfolge sämtliche Regeln der Reihe nach herangezogen werden. Aus diesen wurden dann die anwendbaren auserwählt und durch ihre Anwendung wurde dann die neue Zeichenfolge, welche als Grundlage der Synthese diente, erreicht.

In manchen Beziehungen können die Ergebnisse der Experimentalarbeit als allgemeingültig bezeichnet werden, doch es gibt auch Ergebnisse, welche ausschliesslich für den Versuchstoff gelten. So z.B. ist man in der syntaktischen Synthese von der Voraussetzung ausgegangen, dass sowohl im Englischen als auch im Deutschen der Aufbau der meisten nominalen Ausdrücke der gleiche ist. In den geprüften Text werden die vorausgehenden Adjektiva (Attribute der Hauptwörter) übernommen. Ein wesentlicher Unterschied besteht aber in der allgemeinen Stelle der Satzteile. Es muss die richtige deutsche Wortfolge festgestellt werden. Vorläufig wurde zum Versuch eine für einen jeden Satztypus gültige Wortfolge aufgestellt.

Der sich auf die syntaktische Analyse und syntaktische Synthese beziehende Teil des Stoffes ist sehr wertvoll. Auf diesem Gebiete taucht eine grosse Anzahl von Problemen auf, deren Lösung für die automatische Sprachübersetzung unerlässlich ist. Die Arbeitsgruppe hat eine grosse Anzahl von wesentlichen Fragen gelöst. Im Verlauf der Versuche bestand der längste Satz aus 63 Wörtern. Bereits die syntaktische Prüfung eines so langen Satzes birgt viele Schwierigkeiten in sich. In einem Satz von dieser Länge bedeutet bereits die Festsetzung der Abhängigkeitsverhältnisse eine grosse Schwierigkeit, da voraussichtlich die mehrfache Anwendung von wesentlich mehr Regeln notwendig ist, wie in einem kürzeren Satz. Andererseits sind die zu je einer Regel gehörenden Glieder im Satz ziemlich verstreut.

Die morphologische Synthese ist bereits von geringerem Interesse, da sie zum Teil mechanisch gelöst werden kann. Daher kommen hier nicht so viele grundsätzliche Probleme zum Vorschein, und besonders keine Probleme allgemeiner Natur. Die Probleme sind nicht allgemein, da die morphologischen Probleme der deutschen Sprache und anderer Sprachen voneinander in einem grösseren Ausmass abreichen, wie dies bei den im Verlauf der syntaktischen Untersuchungen auftauchenden Problemen der Fall war.

Was sich auf das englische Analysenwörterbuch und das deutsche Synthesenwörterbuch bezieht, bedeutet auch wenig für uns. Die an die Analysen- und Synthesenwörterbücher gestellte Forderungen sowie der Aufbau von Wörterbüchern sind im hohen Grade von den syntaktisch-morphologischen Untersuchungen, deren Methoden und der bestimmten Sprache abhängig.

Was nun die in der Abhandlung beschriebene Arbeit im grossen und ganzen betrifft, so ist sie für einen jeden von Nutzen, der von einer beliebigen Sprache ins Deutsche übersetzt, da die deutsche Synthese übernommen werden kann. Für die Analyse der Ausgangssprache wiederum bedeutet die für das Englische ausgearbeitete Analyse eine grosse Hilfe.

Um die Übersetzungsarbeit auf einer Maschine vornehmen zu können, muss die Sprache in einer exakten Form beschrieben und die sprachlichen Erscheinungen formalisiert werden können. Daher werden im Verlauf der automatischen Übersetzung die Ergebnisse der mathematischen Linguistik verwendet. Die mathematische Linguistik untersucht die Sprache von mehreren Gesichtspunkten aus und stellt dadurch verschiedene Sprachmodelle auf. Die Abfassung der sprachlichen Modelle ist formal. Dadurch eignen sie sich für die automatische Übersetzung. Die oben behandelte Arbeit bietet ein Beispiel für die Verwendung eines formalen Sprachmodells, zur automatischen in gegebenen Fall englisch-deutschen Übersetzung.

Besprochen von Maria Stein