https://doi.org/10.3176/lu.1968.3.08

ФЕРЕНЦ ПАПП (Дебрецен)

О ПЛАНЕ МАШИННОЙ ОБРАБОТКИ ЭТИМОЛОГИЧЕСКОГО СЛОВАРЯ ВЕНГЕРСКОГО ЯЗЫКА

В течение 1964—1967 гг. нами был обработан на перфокартных (счетно-аналитических) машинах толковый словарь венгерского языка (A magyar nyelv értelmező szótára I—VII, Budapest 1959—1962; всего около 60 000 слов). О результатах этой работы сообщалось и в советской печати.[1] После того, как эта работа была в основном завершена (предстоят некоторые дальнейшие переупорядочения, необходимость которых выявилась в ходе обработки материала), мы решили приступить к аналогичной обработке нового этимологического словаря венгерского языка, первый том которого вышел в свет в конце 1967 г.[2] Ниже постараемся ознакомить читателя с планом машинной обработки этого словаря, обходя технические вопросы и ориентируясь на читателя — лингвиста-финно-угроведа.

I

«Машинная обработка» словаря означает, что мы вносим в память машины каждое заглавное слово его, снабженное лексикографическими сьеденияmии; после этого машина с молниеносной быстротой и сверхчеловеческой (хотя и не абсолютной) точностью способна произвести различные переупорядочения этого материала по внесенным критериям.

Для большей ясности поясним сказанное некоторыми примерами. При обработке толкового словаря для каждого слова была зафиксирована, например, категория «часть речи» — согласно определению толкового словаря. Одно из возможных переупорядочений материала: сгруппировать слова по частям речи, внутри каждой части речи — по алфавиту. В результате мы получили от машин те же 60 тысяч заглавных слов, но так, что сначала следовали глаголы в алфавитном порядке (их оказалось около 10 тыс.), потом — существительные (более 30 тыс.), затем — прилагательные и т. д. Ясно, что для грамматиста очень полезно видеть вместе такое большое количество слов той или иной части речи. Притом машина выявила также некоторые более редкие группы в категории «часть речи». Так, общеизвестно, что в нашем языке довольно широко распространено явление конверсии, которое заключается в том, что одно и то же слово может выступать в роли то одной, то другой части речи. Особенно многочисленны существитель-

[1] Ф. Папп, О машинной обработке одноязычных словарей (На материале венгерского словаря). — Научно-техническая информация (в печати).

[2] A magyar nyelv történeti-etimológiai szótára I, Budapest 1967.

ные-прилагательные и прилагательные-существительные; порою даже кажется, что для нашего языка категории «существительное» и «прилагательное» неадекватны и вместо них надо было бы иметь единую категорию «имени». Это было хорошо известно и без машинной помощи — после указанного упорядочения же мы получили возможность рассматривать в совокупности все слова данного словаря, охарактеризованные составителями как существительные-прилагательные, и все слова, охарактеризованные как прилагательные-существительные. Однако оставалось неясно: а) какие еще двойные комбинации частей речи возможны, по крайней мере в рамках рассматриваемого словаря; б) какие тройные комбинации возможны; в) если такие случаи, когда одно и то же слово может выступать в роли одной из четырех частей речи. Все эти более редкие группы были выявлены, их элементы перечислены (в частности, оказалось, что на самом деле есть восемь заглавных слов с четырьмя характеристиками по категории «часть речи»).

Другой пример из выполненной обработки толкового словаря. При каждом слове было зафиксировано, сколько значений ему приписано составителями словаря. В результате упорядочения по критерию «количество значений» мы имеем полные списки слов с одним значением (как бы «термины»), с двумя значениями, с тремя и т. д., вплоть до многозначных слов. (Наиболее многозначным оказалось в нашем словаре венгерское слово *is* 'тоже' — ему было приписано 101 значение!).

Разные критерии можно комбинировать между собой. Так, относительно двух данных примеров: можно было потребовать списки по частям речи, внутри этого — по количеству значений, далее — слова в обычном алфавитном порядке. После такого упорядочения материала сначала идут глаголы (внутри глаголов: однозначные, потом двузначные и т. д.); затем — существительные (с таким же внутренним подразделением) и т. д. Здесь уже не только списки, но и статистические сведения оказались небезынтересными: выяснилось, что наибольшее количество однозначных слов — среди существительных, более многозначны — прилагательные, еще более многозначны — глаголы и т. д.

Следовательно, «обработка на машинах» значит: неограниченные возможности комбинаций, переупорядочений внесенного материала; получение списков слов, иначе трудно составляемых и т. д. Возвращаясь к нашему плану: самое существенное — определить, что мы должны внести в память машины, так как все возможные дальнейшие упорядочения мыслимы только на основе комбинаций внесенных данных.

II

Вопрос о том, что внести в память машины из материала этимологического словаря, можно разделить на два подвопроса: какие слова внести и какие виды информации зафиксировать при каждом внесенном слове.

А. Первый подвопрос решается сравнительно просто. Мы условились включить в обработку каждое заглавное и каждое производное слово, снабженное внутри словарной статьи необходимыми лексикографическими данными (например, дата первого появления, часть речи и т. д.). Таким образом в первом томе мы будем иметь около 8300 единиц (слов); подготовка этого материала проводится как раз сейчас, во время написания этих строк. А в целом в словаре, запланированном на три тома, мы будем иметь около 30 тысяч слов.

Б. При каждом из этих 30 тысяч слов будут зафиксированы следующие данные:

1. Современные сведения о слове:

а) омонимия (согласно обозначению этого явления в этимологическом словаре);

б) гармония гласных (велярный — палатальный ряд);

в) количество корней, входящих в данное слово (один корень, приставка + один корень, два корня, три корня и т. д. — среди заглавных слов и производных слов нашего источника есть и такие, более сложные единицы);

г) наличие или отсутствие суффикса, тип суффикса;

д) часть речи;

е) является ли данное слово представителем корня или нет?

«Современность» этих данных в том, что они берутся на основе современного нам состояния языка, если слово употребляется и в наши дни. Немало устарелых, полностью выпавших из языка заглавных и производных слов есть и в нашем словаре: в таком случае мы фиксируем эти же виды информации в той форме, что наблюдалась при жизни слова.

Большинство данных (омонимия, часть речи и т. д.), как нам кажется, не нуждается в объяснениях. Хочется сделать замечания лишь относительно двух пунктов.

Гармония гласных (б) фиксируется нами на основе того, какие алломорфы присоединяются к данному корню (в случае сложного слова учитывается по этой же причине только последний корень — качество гласных в грамматических алломорфах определяется всегда этим последним элементом). Отмечаются случаи, когда кодирующий студент колеблется в определении гармонии гласных (например, *buldózerek* или *buldózerok* 'бульдозеры'). Кроме того, отмечается, можно ли однозначно решить гармонию гласных на основе фонемного состава корня. Гармонию гласных можно решить, вообще говоря, если в корне есть одна из фонем *a*, *o*, *u* (велярный ряд) или *ö*, *ü* (палатальный ряд), и нельзя решить, если в корне только фонемы *e*, *i* (и соответственно в каждой из трех групп — долгие фонемы этого же качества). Уже в ходе кодирования (без всякой машинной помощи) были выявлены в связи с этим некоторые интересные нестандартные случаи. Так, слову *костюм* в венгерском языке соответствует слово *kosztüm*. Так как в этом корне есть фонема *o*, слово должно было бы быть велярного ряда, оно же палатального ряда (дат. пад. *kosztümnek*, вин. пад. *kosztümöt*, прилагательное, образованное от этого существительного: *kosztümös* и т. д. — во всех этих и остальных формах гласный в грамматических алломорфах без колебания палатального ряда).

Чрезвычайно труден и в то же время имеет кардинальное значение вопрос о том, является ли данное слово представителем корня (е). Наша цель: после машинной обработки иметь перед собой в одном общем списке все корни, отраженные в нашем источнике, и каждый корень — только по одному разу. Этот общий список будет интересен уже сам по себе: неплохо знать, сколько корней в нашем языке и каковы они. Кроме того, этот общий список можно будет подвергнуть разным дальнейшим переупорядочениям, например по этимологическим пластам. (Можно подумать и о более глубоких подразделениях. Так, можно потребовать от машин, чтобы они составили фонемную статистику по этимологическим пластам корневых слов — какой фонемный

состав и с какой нагрузкой для каждой фонемы характерен для наших корней финно-угорского, тюркского, славянского и т. д. происхождения). Ясно, что производные слова, включенные в нашу обработку, автоматически исключаются: они не являются представителями корней. Однако они «автоматически» выключаются нами, кодирующими людьми: если особо не отметить того, что они — безусловно не корневые слова, машина впоследствии, естественно, никак не может выполнить эту нехитрую операцию. Но есть гораздо более сложные случаи. Очень часто не известна этимология того или иного слова, есть только разные гипотезы. Если принять одну из гипотез — то оказывается, это — особый корень, если принять другую — то оказывается, что данное слово является производным от другого, при котором нами уже отмечено, что оно представляет собой особый корень. В других случаях мы имеем дело с более или менее старинными случаями словосложения. Так, венг. *arc* 'лицо, щеки' — результат старинного словосложения, происшедшего еще в общую финно-угорскую или в угорскую эпоху. Элементы этого словосложения *orr* 'нос' и *száj* 'рот' — и ныне живущие венгерские слова, они, естественно, фигурируют как представители особых корней. Что же сказать о самом слове *arc* (бывшее *orr* + *száj*)? Представляет оно собой особый корень для венгерского языка или же сложное слово и как таковое не может быть особым корнем? (Вследствие чрезвычайной давности предполагаемого словосложения *orr* + *száj*, конечно, ни один из венгров—нефилологов не подозревает, что *arc* — сложное слово). Или, что делать с венгерскими словами *gitár* 'гитара' и *citera* 'цитра'? В конечном счете оба они восходят к одному и тому же (греческому) этимону, но попали в венгерский язык через разные каналы и в настоящее время опять-таки не ощущается между ними никакой связи, т. е. это один из случаев довольно частого явления параллельных заимствований. Как здесь следует поступить? Кодирующему предоставлена возможность выбрать в случае каждого слова между восемью разными ответами на вопрос о том, представляет ли данное слово корень или нет. Ответы могут быть такого типа: слово — безусловно представитель особого корня, слово — представитель особого корня только внутри венгерского языка, слово — может быть, не является представителем особого корня, слово — вероятно, не является представителем особого корня и т. д. Я думаю, уже сами списки с «безусловными корнями», «только венгерскими корнями», «вероятными корнями», «вероятными некорнями» и т. д. без всяких дальнейших комбинаций представят интерес для лингвистов.

2. Лексикографические данные:

а) количество вариантов, примеров употребления и т. п. данного слова внутри словарной статьи;

б) количество значений (общее количество и количество и в настоящее время употребляемых значений);

в) в случае непроизводных слов: количество производных от него внутри данной словарной статьи.

3. Исторические данные:

а) абсолютно первое появление данного слова (дата, часть речи);

б) первое настоящее появление данного слова (дата, часть речи);

в) предполагаемая эпоха, когда данное слово на самом деле появилось в нашем языке;

г) уточнения к датам абсолютно первого и самого первого появления (это — модификации вроде: «после», «перед», «около» и т. п.).

«Абсолютно первое появление» — это дата, когда данное слово

появилось в какой-либо форме в одном из наших памятников. Это может быть имя собственное (имя лица, топоним в грамотах), производная форма и т. п. Возможно, что первое появление этого слова и есть его «настоящее» появление, т. е. та форма, в которой оно существовало дальше и живет, быть может, сейчас. Например, для венг. *árpa* 'ячмень' абсолютно первым появлением является дата ок. 950 г. н. э.: в одной из греческих исторических работ это слово впервые зафиксировано как часть собственного имени (имени вождя). А первое настоящее появление его относится только к 1395 г., когда оно в этой своей исходной форме встречается в одном из венгерских письменных памятников. Часть речи как в случае абсолютно первого, так и в случае самого первого появления может быть другой и данные могут отличаться от принадлежности к части речи этого слова в современном состоянии языка: на основе трех (или двух, если абсолютно первого появления не было) разных сведений, взятых из трех (двух) синхронных срезов языка, можно будет сделать интересные выводы.

Кроме первого появления слова в письменных памятниках, мы посчитали необходимым зафиксировать также и эпоху предполагаемого появления данного слова в нашем языке, так как первое письменное появление того или иного слова может быть случайным. Если слово не относится к религиозной или юридической терминологии наших самых ранних грамот — оно может появиться сравнительно поздно, несмотря на то что, безусловно, уже существовало и раньше. Так, слово древнего финно-угорского корня *apa* 'отец' появляется почему-то впервые только около 1395 г., упомянутое *arc* 'лицо, щеки' — в памятнике, датируемом позже 1372 г. Ясно, что эти даты в некотором смысле нереальны. Легко будет потребовать от машин списки специально отобранных слов, в случае которых наблюдается слишком большое расхождение между их первым появлением и предполагаемой эпохой их появления в нашем языке. Эти слова надо будет просто специально искать в еще не обработанных памятниках; в своем большинстве они, очевидно, и будут найдены. Сходным образом будут выявлены случаи, когда слишком большое временное расстояние разделяет абсолютно первое и самое (настоящее) первое появление того или иного слова: здесь, возможно, и будут обнаружены случаи, когда отождествление «абсолютно первого» и «самого первого» появления было фальшивым, когда, следовательно, речь идет не об одном слове, а о двух. Преимущество машинной обработки опять проявляется в том, что лингвист получит возможность видеть в одном списке то, что раньше было разбросано по словам. Эти совместные списки могут натолкнуть исследователя на интересные мысли.

4. Этимология:

а) определение языка-передатчика;

б) определение языка-источника;

в) уточнения к (а) и (б) типа «вероятно», «может быть» и т. п.

Как язык-передатчик, так и язык-источник определяется в двух «тактах». В первом фиксируется как бы главное направление происхождения: финно-угорское, собственно венгерское, тюркское и т. п. Во втором это направление уточняется. Так, в частности, внутри группы «финно-угорское» выделяются подгруппы: общефинно-угорское, общеугорское (т. е. венгерское — хантыйское — мансийское). Более того: кроме «общеугорских» слов особо отмечаются слова, находящиеся только в венгерском и хантыйском языках, только в венгерском и ман-

сийском (т. е. как бы «общехантыйско-венгерские», «общемансийско-венгерские»); особо выделяются слова, имеющиеся только в венгерском и пермских языках (т. е. «общевенгерско-пермские»). Внутри общей большой группы «тюркские заимствования» выделяется несколько подгрупп (тюркское до занятия Карпатского бассейна венграми, турецкое и т. д.). Вследствие географического и культурного расположения венгерского народа с особой тщательностью разработана общая группа «немецкие заимствования» (особая большая группа, оторванная от «германских языков» вообще), а также группа «славянские заимствования», так как венгры непосредственно соприкасались со всеми тремя группами славянства в разные периоды их жизни. Надо ли говорить о том, сколь интересно будет видеть в одном списке, например, все «хантыйско-венгерские» слова или все слова «либо словацкого, либо южнославянского происхождения». Как известно, в словацком языке немало южнославянизмов — по сей день не совсем ясна причина этого: виноваты ли здесь венгры, оторвавшие предков словаков от их ближайших родственников — южных славян или она кроется в чем-либо еще. Производные слова могут быть рассмотрены особо, поэтому можно будет уделить внимание и такому немаловажному вопросу, как: с какой продуктивностью участвуют разные этимологические пласты в образовании новых венгерских слов? Другими словами, сколько производных слов образовано от корней финно-угорского, тюркского, славянского и т. д. происхождения?

5. Иные данные:

а) стиль (только с той точки зрения, употребляется ли данное слово и в настоящее время в литературном языке, или оно является устарелым, диалектальным и т. п.);

б) этимологическая характеристика по этимологическому словарю Г. Барци[3];

в) разное (обращается внимание исследователя на какую-нибудь интересную черту этимологии, истории и т. д. данного слова).

III

Как проводится подготовка к машинной обработке? Об этом вопросе можно говорить, оставаясь в сфере лингвистики: эта подготовка — дело лингвиста, хотя она уже и слегка соприкасается с сугубо техническими вопросами.

После определения обрабатываемого материала (2) делаются бланки для кодировщиков. Каждый бланк содержит несколько строк — в нашем случае — 10: в каждой строке будет по одному слову. Бланк распределен не только горизонтально (по строкам), но и вертикально; строки разбиты на столбцы. В столбцах последовательно помещаются данные раздела: в первом столбце само кодируемое слово, потом — знак омонимии, количество корней, суффикс и т. д.; даты первых появлений; этимологии языка-передатчика и языка-источника и т. д.

Сначала эти бланки пустуют. В качестве самого первого этапа машинистка берет этимологический словарь и перепечатывает оттуда на бланки, в столбец, предоставленный словам, каждое заглавное и производное слово. После этого бланки со словами передаются кодировщикам — в нашем случае это студенты-добровольцы, специалисты по истории, венгерскому языку, славянской филологии. Согласно нашим инструкциям, они впоследствии и заполняют еще пустующие столбцы на бланках при каждом слове. Так, в столбце, предназначенном для

[3] G. Bárczi, Magyar Szófejtő Szótár, Budapest 1941.

определения «количества корней», они пишут цифру 1, если слово состоит из одного корня, 2, если оно состоит из двух корней и т. д.; в столбце, предназначенном для даты первого появления, — цифры даты первого появления; в столбце, предназначенном для этимологии, — цифры: 1 — если слово (корень) финно-угорского происхождения, 2 — если оно собственно венгерское, 3 — если оно тюркское, 4 — если оно славянское и т. д. (т. е. даже эти цифры довольно понятны: мы идем в основном в хронологическом порядке этимологических пластов).

Приняты всевозможные меры предосторожности. Так, фактически каждый бланк заполняется в двух экземплярах; впоследствии эти экземпляры сравниваются и ошибки исправляются. Приняты и некоторые другие меры такого рода. А затем — наше дело закончено. Весь материал передается специалистам по машинам, и их действия могут уже нас не интересовать.

IV

Нас интересуют только те результаты, которых мы ждем от обработки подготовленного нами материала на машинах. В ходе изложения закодированной информации я уже не мог удержаться и указывал мимоходом на некоторые ожидаемые списки по корневым словам — некорневым словам, по этимологическим пластам и подпластам и т. д.

Легко представить себе дальнейшие подобные списки. Так, мы будем располагать «хронологическим словарем» венгерского языка: мы будем иметь списки слов, появившихся в ту или иную эпоху, списки слов, которыми располагал наш язык в то или иное время и т. д. Критерии эпохи, этимологии, части речи и т. д. могут быть скомбинированы между собой; наряду с полными списками можно получить краткие, сжатые статистические таблицы относительно всех этих данных или других, внесенных нами в память машины. К тому же все это делается на очень нехитрых, старых (им скоро 100 лет) электромеханических, счетно-аналитических машинах, которые обязательно есть в каждом сколько-нибудь значительном городе Советского Союза. Лингвисты только не знали до сих пор об этих машинах — на них ведут бухгалтерский учет, инвентаризацию, работы по статистике и планированию народного хозяйства.

«Традиционные» лингвисты нередко критиковались как позитивисты, которые не видят системности в диахронии. Да, действительно, системность в диахронии трудно уловить: можно хорошо написать одну этимологию, десяток этимологий, можно даже собрать сотни и тысячи этимологий в одном этимологическом словаре — но это еще не система. В ходе отдельных этимологий исследователи нередко делали ссылки на аналогичные явления, на аналогичное изменение смысла в других случаях — но и это еще не было системой, это были эпизоды исследовательской работы. Нам кажется, что использование машин и в этой области может оказать значительную помощь. Мы будем иметь возможность видеть вместе то, что до сих пор знали только по деталям — это может послужить толчком для обнаружения системности и там, где она теоретически предполагалась и раньше, но практически ее трудно было найти. Все это, конечно, толчок, помощь и т. п.; пока мы не собираемся поручить машинам решать исследовательские лингвистические задачи, хотя, конечно, и это не невозможно.[4]

---

[4] Ср.: И. А. Мельчук, Автоматизация в лингвистике. — ВЯ 1968, № 1, стр. 138—144.

FERENC PAPP (Debrecen)

ON A PLAN FOR THE MECHANICAL PROCESSING OF THE ETYMOLOGICAL DICTIONARY OF HUNGARIAN

After the processing of the Explanatory Dictionary of Hungarian by the Mathematical and Applied Linguistics Group in Debrecen, performed with the aid of electromechanical and electronic machines, the same group has planned to process the Etymological Dictionary of Hungarian by similar methods. Every entry word appearing in the dictionary will be included, together with every derived word listed. The items of interest in the processing include the synchronic characterization of the words from the viewpoint of the modern language, lexicographic data, history (date of first occurrence, original part of speech, etc.), etymology (immediate and ultimate source language), and other questions (stylistic value, vowel-harmony class, etc.). Present status of the work: the 8400 words occurring in volume I (A — Gy) of the dictionary which appeared in the autumn of 1967, have been encoded twice for purposes of checking; after the comparison and correction of the code sheets, the material will be prepared for electromechanical processing by the end of 1968 or the beginning of 1969.