Применение ЭВМ в изучении лексики венгерского языка

ФЕРЕНЦ ПАПП—МИХАЙ ФЮРЕДИ

(Дебрецен—Будапешт)

Четверть века назад были предприняты первые попытки применить современные ЭВМ в изучении венгерского языка. Эти эксперименты касались фонемной статистики, и в дальнейшем продолжались работы в этом направлении. Вскоре после этого начались и ведутся сейчас исследования в венгерской морфологии: автоматический синтез, далее анализ венгерских словоформ, в последнее время — автоматический венгерский синтаксический анализ. Комплексные работы по МП с венгерского на русский и обратно за границей начались уже в середине 50-ых годов. Изучение размера венгерского стиха, а в самое последнее время — использование микрокомпьютеров в преподавании и отчасти изучении венгерского языка: вот дальнейшие вехи в этом длинном и сложном деле.

Ниже мы намерены изложить только некоторые результаты и наметить некоторые планы в области, где сходные работы начались тоже сравнительно рано, в начале 60-ых годов: в области изучения венгерской лексики. Здесь, видимо, возможны два подхода: исходя из текстов составлять их словари, с одной стороны, и исходя из уже готовых словарей, обработать их материал, с другой стороны. Сначала мы изложим некоторые вопросы первого подхода в связи с машинной обработкой Этимологически-исторического словаря венгерского языка (**1.**) и второго в связи с подготовкой «большого академического словаря» венгерского языка («Nagyszótár») (**2.**), а также — в связи с первым разделом частотного словаря венгерского языка (**3.**).

1. О плане машинной обработки Этимологически-исторического словаря венгерского языка (ЭИС) было сообщено раньше (см., напр., Советское финно-угроведение 4: 207—214). Теперь мы располагаем некоторыми данными относительно 1-го тома ЭИС. Это — примерно одна треть всего материала, 8 с лишним тыс. первичных и вторичных заглавных слов и производных лексем, помещенных в словарных статьях. В качестве примера общих, глобальных вопросов, которые можно поставить перед машинной обработкой, сошлемся тут только на два, надеясь, в их случае не очень мешает обстоятельство, что это — лишь треть материала, к тому же не случайно выбранная: начало алфавита, от A до Gy.

а) Если посмотреть на таблицу № 1, мы видим применяемую нами хронологизацию, а также количественные соотношения появления лексем-корней в каждую эпоху. «Появление лексем» — это, естественно, весьма условное понятие: все зависит от того, что найдено и что не найдено; какие памятники обработаны и какие — нет. Семь лексем-корней, встречаемых в ранних греческих грамотах, и без машинной помощи известны специалистам

(*álom*, 'сон', *álmos* личное имя, *árpa* 'ячмень', *búcsú* 'разрешение' [позже: 'паломничество' и др.], *csorog* 'течь', *fal*[1] 'жрать' и *gyula* ('воевода'); они скорее иллюстрируют, как мы здесь понимали лексему-корень и ее первое появление.

Или возьмем другой «конец» этого списка: 44 лексемы-корня, появившихся в венгерском после освобождения 1945 г. и уже попавшие в ЭИС, хотя ясно, что первоначальным профилем таких словарей не является собирание новейших элементов. Вопреки ожиданию — там нет ни одного русского элемента в узком смысле слова! Главная причина, кажется, в том, что ожи-

Табл. 1

Эпохи первых письменных фиксаций лексем-корней в т. 1 ЭИС

Эпоха		Количество лексем-корней		100% = 8278*	
I: др-в	1) до 960 г.	7	2492	*30*	
	2) до 1191 г.	205			
	3) до 1526 г.	2280			
II: ср-в	4) до 1771 г.	2316	2316	*28*	5486 *(66)*
III. нов-в	5) до 1899 г.	3170	3460	*42*	300 *(4)*
	6) до 1944 г.	256			
	7) до 1960 г.	44			

* По данным составителей ЭИС таких элементов в т. 1 должно быть всего 8386. Мы сейчас ищем потерявшиеся у нас элементы. Сокращения: др-в — древневенгерский и т. д.

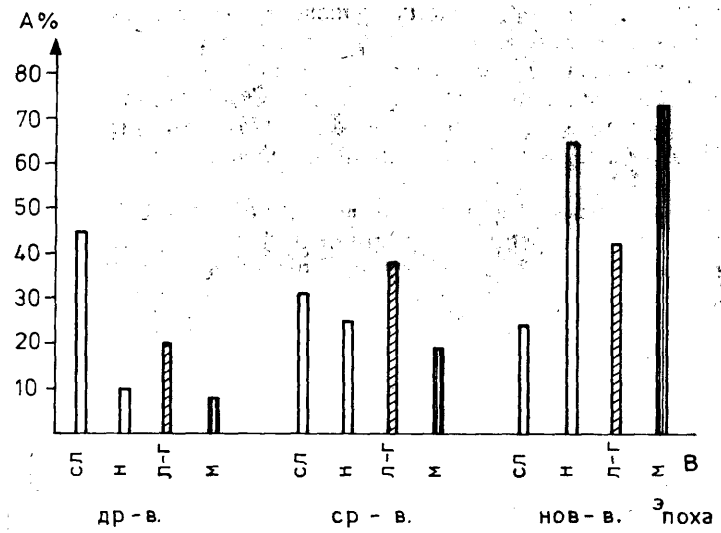
даемые заимствования попали в венгерский язык раньше: *bolsevik* 1907, *bolsevizmus* 1917, *bolsevista* 1919; из других томов: *kolhoz* 1934, *szovjet* 1917 и т. д. Нет, язык нелегко раскрывается: великие исторические перемены не отражаются в нем так поверхностно. Если лучше присмотреться, то оказывается — те международные слова, общие европеизмы, новые технические выражения и т. п., которые раньше веками попадали в словарь с запада — теперь вдруг приходят с востока, именно из русского языка. Вот элементы, о которых составители словаря пишут[3], что «под русским и немецким влиянием», «по всей видимости, непосредственно из русского» и т. п.: *agresszor*, *aktivizál*, *aspirantúra*, *centrizmus*, *centrista*, *diszpécser*, *diverzáns*, *fasizál*, *fesztivál*, *fetisizál*. Или еще глубже и труднее уловимо: о лексеме *egyenlősdi* 'уравниловка' пишется только: «совсем нового происхождения» и единственный источник: русско—венгерский словарь Хадровича и Гальди 1951 г. Это не калька, просто надо было найти название самому понятию «уравниловка», независимо от языка, в данном случае — русского; само понятие стало все чаще появляться.

Ясно, что среди этих 44 новейших элементов есть и другого происхождения: *dauerol* 1948 'делать перманент' (вид прически) — немецкого, упомя-

нутое *egyenlősdi* собственно венгерского, таковы же из «собственных» или ранее заимствованных элементов сделанные разговорные новообразования: *dagi* 'толстушка', *focista* 'футболист' и т. п.

б) Сказанное выше о прямых и более глубоких реакциях языка на исторические изменения становится еще более наглядным, если посмотреть на рис. 1. Там эпохи появления лексем комбинируются с историей-этимоло-

Рис. 1. Предварительное распределение появления лексем разных этимологических пластов по т. 1 ЭИС



Объяснения к рис.: А % соответствующего этимологического пласта впервые появилось в эпоху В (100 = общее количество лексем-корней данного этимологического пласта: сл[авянский], н[емецкий], л[атинско]-г[реческий], м[еждународный].

гией этих же элементов. Видно, что, попав в Карпатский бассейн, сначала надо было «освоиться». По свидетельству 1-го тома почти половина всех славянских заимствований попала в наш словарь очень рано: до 1526 г. Потом заимствований из этой среды становится все меньше и меньше. Но зато, параллельно с этим, растет количество заимствований из латинского и греческого языков, из немецкого языка и, прежде всего — растет пропорция интернационализмов: почти три четверти последних элементов попали в наш лексикон за последние два столетия. Мы бы сказали: освоившись в Карпатском бассейне — венгры должны были найти свое место в Европе, связаться с Европой и путем слов.

2. Проект о создании нового «большого академического» словаря (БС) был одобрен президиумом АН Венгрии 28.2.84 г. Согласно этому проекту БС должен быть машинной сокровищницей венгерского литературного языка от начала книгопечатания в Венгрии до наших дней, общим объёмом в 10 млн. словоупотреблений. В самом общем виде об этом проекте пока можно сказать следующее. Готовятся не карточки (таких карточек по старому плану хра-

нится в Институте языкознания АН Венгрии несколько миллионов), а фиксируются полностью разные тексты небольшими отрывками (по 1 авторскому листу каждый). Из этих текстов готовятся частные конкордансы — и один общий, для всего материала. Соответствующие строчки конкорданса и представляют собой «карточки»; с них, если это понадобится, и можно печатать БС традиционного (т. е. печатного) типа.

В тесном контакте с проектом БС готовились и готовятся конкордансы по отдельным выдающимся писателям, поэтам. Соответствующая доля из их произведений вольется в общий фонд БС — но независимо от этого будет доступен и частный конкорданс к их произведениям. К лету 1984 г. изготовлен конкорданс ко всем произведениям, письмам, переводам Б. Балашши и ко всем стихотворениям Э. Ади; готовятся конкордансы к произведениям М. Чоконаи-Витеза и А. Йожефа. Вне рамок БС, но сходным образом в Дебреценском университете была произведена сходная машинная обработка двух кодексов середины XV в.: кодекса Йокаи и кодекса Бирк.

3. По текстам же современного венгерского языка на основе более полумиллиона словоупотреблений составлен частотный словарь венгерской художественной прозы (ЧС). Работы начались десять лет тому назад. Ручное расписывание слов по первому жанру (это и есть художественная проза) закончилось в 1976-ом году, но машинная обработка последовала с некоторым опозданием, лишь в 1979-ом году. При участии нескольких сотен студентов, использованием нескольких десятков часов машинного времени к 1983-ему году окончились все работы по первому жанру, и весь материал готов к изданию.

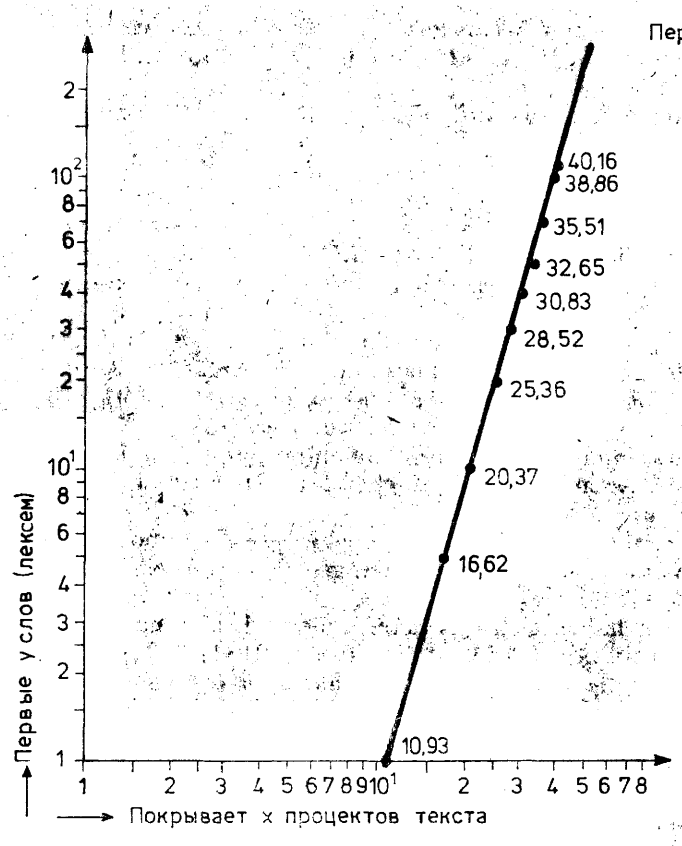
Может быть излишне говорить, сколько усилий потребовала от редакторов и сотрудников этого словаря стандартизация мнения студентов из каждого уголка страны. Надо признаться, что в осуществлении плана ЧС пришлось намного больше полагаться на человеческий ум и руки, чем это было бы целесообразно в подобном деле. Несмотря на это мы можем сказать, что этот ЧС заслуживает внимания не только лингвистов финно-угроведов, но и специалистов по квантитативной лингвистике и типологии. В ЧС обработаны и лексемы, и словоформы. Список по убыванию модифицированной частоты (о модифицированной частоте см. A. Juilland—E. Chang-Rodriguez: Frequency Dictionary of Spanish Words. London — The Hague—Paris, 1964 стр. LXVII и сл.) составлен на лексемы, и под каждой лексемой приведены все встретившиеся словоформы с дисперсионными и частотными данными (тоже по убыванию модифицированной частоты). Алфавитный список тоже содержит все данные по лексемам, и под каждой лексемой приведены все словоформы по алфавитному порядку. Собственные имена (простые и сложные) исключены из основного списка. Они будут обработаны отдельно. Редакторы ЧС Й. Келемен и М. Фюреди не следовали за принятой многими практикой исключения и числительных: они фигурируют в обоих основных списках и составляют 1,7% всех словопотреблений. Для обеспечения дальнейшей машинной обработки записаны на магнитные пленки ЭВМ и устойчивые словосочетания и фразеологизмы.

До сих пор мы говорили о полном («машинном») варианте ЧС. Из всего материала будет опубликовано лишь 3500 лексем со своими словоформами по убыванию модифицированной частоты; алфавитный список будет включать в себя лексемы с частотой $Ч_{мод} < 5{,}00$ (5684 лексемы, без указания на слово-

формы, с отметкой, если данная лексема фигурирует в частотном списке подробно); будут приложены и некоторые статистические таблицы о структуре словаря, о встречаемости частей речи, грамматических категорий.

В составлении текстов выборки редакторы стремились учесть и лингвистические, и статистические требования. Все 508008 словоформ ЧС происходят из 258 разных отрывков текста, причем ни один из авторов не представлен более чем тремя произведениями (и даже в случае двух или трех произведений потребовалось, чтобы они были взяты из разных произведений данного писателя). Этим удалось снизить возможность чрезмерного влияния одного или нескольких авторов. Предпринимались и дальнейшие шаги в этом направлении: вычислены величины дисперсии и модифицированной частоты внутри одного и того же жанра (при случайном разбросе и распределении материала на пять одинаковых по величине частей). Т. е. результаты еще больше сглаживались, и окончательный частотный список — выражаясь термином математической статистики — более надежен.

В качестве иллюстративного материала приводим рисунок, на котором показывается покрываемость текста начала частотного списка лексем (см. Рис. 2).



Первые y лексем покрывает x % текста

y	x %
1	10,93
5	16,62
10	20,37
15	23,22
20	25,36
25	27,05
30	28,52
35	29,76
40	30,83
45	31,77
50	32,65
60	34,15
70	35,51
80	36,78
90	37,88
100	38,86
110	39,75
115	40,16

Рис. 2