https://doi.org/10.3176/lu.2000.2.08

МАРГИТ ЛАНГЕМЕТС (Таллинн)

СЛОВАРИ ЭСТОНСКОГО ЯЗЫКА В ИНТЕРНЕТЕ

В данном обзоре предполагается выяснить, какие словари эстонского языка можно обнаружить в Интернете и в каком виде они там представлены.

KeeleWeb

В 1998 году в Интернете была открыта страница под названием KeeleWeb (*ee.www.ee*)[1], предназначенная по идее для всех материалов по эстонскому языку — словарей, текстовых корпусов, справочников, разного программного обеспечения, в том числе поисковых систем, которые учитывали бы своеобразие эстонского языка. В настоящее время на странице разместились шесть т.н. словарей KeeleWeb, а также отсылки, указывающие местонахождение некоторых других словарей. Словарь KeeleWeb отнюдь не означает, что данный словарь создан только что: давно знакомый нам в форме книги, он помещен в компьютер таким образом, что его можно использовать с помощью Интернета и иным по сравнению с книгой способом. Его можно не только читать подряд, но и пользоваться выборками по запросу, углубляться в языковое явление в пределах одной или нескольких инфокатегорий либо в объеме всего словаря — в зависимости от сложности вопроса, для решения которого построена система запросов. Если задуматься о том, имеет ли вообще смысл применение вычислительной техники в лексикологии, то (наряду с облегчением повседневной работы) именно такой обстоятельно «прочесывающий» запрос, который сортирует материал с учетом разных пожеланий, и представляет собой одно из преимуществ машинного варианта перед книжным. Правда, нового содержания компьютерный вариант не предлагает, но позволяет подойти к имеющемуся под разными углами зрения. Известно, что всем словарям мира присущ общий недостаток: они таят в себе гораздо больше, чем на первый взгляд кажется. Именно поэтому в связи с компьютерными словарями чрезвычайно важное значение имеет обработка имплицитной информации в эксплицитную, превращение невидимого в видимое, что обычно является очень трудоемкой и дорогостоящей работой. Синтаксических данных наши словари, как правило, не содержат (что отнюдь не означает отсутствия в них синтаксических связей), но если присмотреться к примерам-предложениям в отдельных словарных статьях толкового словаря эстонского языка, грамматика начинает заявлять сама о себе: обращаешь внимание на определенные члены предложения, отдельные словоизменительные формы, особенности управления, временные формы и иной контекст. Собранные в словаре образцы предложений составляют идеальный текстовой корпус, который частью исследователей оценивается выше, чем просто корпус, как раз за лексическую мно-

[1] Здесь и далее подчеркивание обозначает поисковый стрит, т.е. текст, который заносится в графу поиска и в Интернете ведет себя как ссылка, направляемая в нужное место. В скобках дан адрес словаря (или иного документа) в Интернете.

гозначность описания (Pustejovsky 1995). Существовавшее доселе мнение теоретиков-языковедов и лингвистов-прикладников о том, что словари особого интереса не представляют, меняется: в 1990-х годах опубликовано несколько лингвистических работ, подтверждающих, что большая часть информации о структуре предложения лучше описывается именно с лексической точки зрения (см., например, Goldberg 1992; 1995).

Тем самым косвенно соединяются и попытки изобрести способы, с помощью которых пояснения, указания и расшифровку условных знаков из начала книги можно перенести непосредственно в словарную статью, поместить рядом с ней или между статьями. Преследуется та же цель: все, что содержится в книге, должно дойти до пользователя. Большую помощь при этом способен оказать компьютер — требующие дополнительных пояснений термины, сокращения, символы передаются в виде (цветных) ссылок: достаточно нажать на них и интересующие пояснения появятся. Ссылки служат и сигналом: в этом месте имеет смысл что-то уточнить.

Система запросов нужна прежде всего исследователям — от узких специалистов-лингвистов и выполняющих курсовую, дипломную или иную работу студентов до школьников.

Компьютерный вариант орфографического словаря Õigekeelsussõnaraamat (*ee.www.ee/QS/*) уже солиден по возрасту, ему 19 лет, и поскольку он предназначен для лингвистических исследований, материал упорядочен и дополнен (по сравнению с ÕS, опубликованным в 1976 году) грамматической информацией (см. Viks 1981 : 211—216). На базе того же автоматического словаря эстонского языка в 1992 году подготовлен и «Краткий морфологический словарь эстонского языка» Юлле Викс (Viks 1992), который позже в свою очередь послужил базой для всей автоматической системы словоизменения эстонского языка (синтез и анализ).[2] В компьютерной версии ÕS все эти дополнения присутствуют, заглавное слово снабжено следующими сведениями: а) часть речи, б) исключения, в) граница сложного слова. Сортировка материала ÕS возможна двумя путями: простым и сложным. Простой запрос действительно прост: задай машине (заглавное) слово, получишь его грамматическое описание. Этот запрос по сути равноценен открытию морфологического словаря на нужной странице (необходимо лишь заучить значения использованных в книге условных знаков) и подходит для тех, кому машина больше по душе, чем книга, или же если книги почему-либо нет под рукой. Сложнее комплексный запрос, позволяющий поисковые признаки комбинировать между собой.

Три словаря — антонимов Antonüümsõnastik (*ee.www.ee/Anton/*), синонимов Sünonüümsõnastik (*ee. www.ee/Symon/*) и фразеологизмов Fraseoloogiasõnaraamat (*ee.www.ee/Fras/*) — представляют собой компьютерные версии «настоящих» книг (Õim 1991; 1993; 1995), что позволяет прочесать словари и по отдельным частям словарной статьи. Словарные статьи варьируют в зависимости от типа словаря, то же касается и содержания их частей. Например, в отношении антонимов заглавным словом является пара антонимов целиком, в словаре синонимов — первый член синонимного ряда, а в фразеологическом словаре — одно или несколько застывших выражений: фразы или сложные слова. В словаре антонимов есть часть словарной статьи, названная «синонимы противоположностей», где отдельно даются синонимы каждой пары. В то же время, например, некоторые из этих синонимов встречаются и в качестве заглавных слов в составе иной пары антонимов, среди синонимов которой обнаруживаются в свою очередь другие слова. На экране компьютера видна паутинка нитей, связывающих слова между собой, причем гораздо отчетливее, чем на страницах книги.

[2] В программном обеспечении Института эстонского языка (*www.eki.ee/tarkvara/*) есть программа морфологического модуля, которая а) шлифует, б) выявляет в исходном форме слова словоизменительный тип и часть речи и в) генерирует слову в исходной форме все эти словоизменительные формы. Как программное обеспечение KeeleWeb (*ee.www.ee*) можно использовать Morfoloogiline analüüs, а как демоварианты Morfoloogiline süntees и Poolitus.

Tesaurus (*ee.www.ee/Tesa/*) тоже предлагает информацию о лексических взаимосвязях между словами, правда, поначалу ограничиваясь лишь синонимами и антонимами. Гипо- и гиперонимы пока отсутствуют, подождем, когда будет готова Eesti WordNet. Не выделены и слова, имеющие формальное сходство: омонимы и паронимы. В значительной мере тезаурус основан на словарях синонимов и антонимов, новое в нем — заглавные слова снабжены знаком части речи. Весь материал для тезауруса эстонского языка построен Философтом (Filosoft — фирма, которая предлагает программное обеспечение в области лингвистики). Вместе с новейшими версиями WORD его можно приобрести непосредственно для своего персонального компьютера и использовать при стилистической обработке текста, обогащая формулировки синонимами и антонимами, предлагаемыми словарем.

Из терминологических словарей на странице имеется англо-эстонский компьютерный словарь Arvutisõnastik (*www.ioc.ee/arvutisonastik/*), который по сравнению с книгой (Hanson, Tavast 1996) в качестве дополнительной функции предлагает поиск по эстонскому пояснению, или соответствию.

Наряду со словарями, KeeleWeb представляет и другие лексические материалы, поначалу в области ономастики. Kohanimed (*ee.www.ee/Kohad/*) — база данных, основанная на атласе Эстонии фирмы Регио (Regio) и созданная в Институте балтийских исследований. KeeleWeb предлагает только топонимы: при желании познакомиться с картами следует обратиться к атласу. С помощью метасимволов и достаточно длинного перечня ограничений (часть города, уезд, улица и т.д.) можно составить вполне приличные вопросы типа «Какие топонимы состоят из двух букв?».

Пользователей KeeleWeb, во всяком случае до сих пор, больше всего привлекает Isikunimed (*ee.www.ee/Nimed/*) — как личные имена, так и фамилии. Материал извлечен из базы данных Регистра по учету населения Эстонии (AS Andmevara) и отражает состояние на осень 1995 года. Сопровождающее предупреждение — «в открытом пользовании — личные имена и фамилии, которые встречаются у населения Эстонии более 5 раз» — очевидно, плохо сформулировано, поскольку часто остается пользователями незамеченным и приводит к неверным толкованиям и упрекам типа «Но меня же так зовут!» и «Но ведь я же существую!». Нельзя по KeeleWeb и искать родственников, опять же по причине конфиденциальности. Имена в KeeleWeb (5509 личных имен и 38 367 фамилий, т.е. те, что встречаются шесть раз и больше) призваны удовлетворять прежде всего научные потребности антропонимики, а также потребительские интересы, но не обслуживать социальные и национальные науки. Для исследовательской работы, требующей более точных данных, необходимо специальное разрешение.

Запрос Ühispäring на странице KeeleWeb ищет информацию сразу во всех названных здесь подсистемах и плюс еще некоторые. На первый взгляд этот запрос может показаться немного бессмысленным: кому нужно, например, сообщение о том, что «лексема *globaliseeruma* среди топонимов не встречается». До сих пор нет ограничения, позволившего бы запросить «все последовательности, которые являются как словами (т.е. встречаются в словарях), так и именами собственными (топонимами или личными)»; антропонимист-исследователь в ответ на такой запрос сразу получил бы много полезной информации разного рода.

Иные словари

Частотный и толковый словарь SASS (ссылка есть и на странице KeeleWeb: *ee.www.ee/filosoft/html_sass/*) анализирует по одному слову целиком документ HTML (по адресу) или непосредственно введенный в окно текст. Жаль, что в SASS отсутствуют более подробные пояснения, что есть что.

Slängisõnastik (*www.eki.ee/dict/slang/*) — это компьютерная версия первого эстонского словаря сленга (Loog 1991), вместе с механизмом поиска и отсылочным

аппаратом к сленгу других языков мира словарь уже некоторое время используется в Интернете. Сленг активен всегда и везде, так и в Интернете: среди словарей доминируют конкурирующие между собой сленговые словники и словнички. В лексикологическом плане они в большинстве своем являются шлаком, что отнюдь не исключает, конечно, возможности для серьезного исследователя найти там для своей работы удачные примеры живой речи.

Murdesõnastik (*www.eki.ee/dict/vms/*) позволяет обращаться с запросами к известному в форме книги словарю эстонских диалектов (Väike murdesõnastik 1982—1988). Своя запись искомого отрывка достаточно сложна для филолога или любого другого человера, далекого от вычислительной техники. В то же время легко сформулировать свой вопрос по приходам, обозначая пожелание с помощью предлагаемых фильтров. Диалектологу особенно по душе возможность связать словарь диалектов с индексом словаря А. Сааресте (Saareste 1979) (*www.eki.ee/dict/saareste/*).

Введен в Интернет и Hargla murraku konsonantism (*www.eki.ee/dict/hargla/*)[3]; это грамматическая база данных, которую Х. Неэтар и Ю. Викс представляют как вспомогательное средство для составления морфологических обзоров диалектных слов и которая в перспективе будет пользоваться спросом при исследовании диалектов. По каждой вводимой словоформе (слову) дается множество лингвистических данных, например, исходная форма слова (в фонетической транскрипции), фраза, частью которой оно является, исходная форма заглавного слова в литературном языке, значение заглавного слова, часть речи, название словоизменительной формы и многое другое.

В связи с KeeleWeb не упомянут еще один двуязычный словарь Inglise-eesti-inglise sõnaraamat (*www.ibs.ee/dict/*), авторские права на который принадлежат Институту балтийских исследований (IBS) и Кинексу (Kinex) и который, согласно своей рекламе, предлагает соответствия более чем 17 000 английских слов. Слова отыскиваются только по определенному (начальному) сочетанию (метазнаки не признаются), в английской части, к сожалению, можно искать только заглавные слова, тогда как выражения внутри словарной статьи могут остаться не обнаруженными. Тем самым английская часть никаких преимуществ перед книжным вариантом не имеет: просто открой книгу в нужном месте и читай! В эстонской части картина несколько иная, но вследствие не столько продуманности, сколько случайности: эстонский является целевым языком словаря, а потому языком соответствий, т.е. присутствует во всех словарных статьях. В результате картина здесь гораздо полнее.

Кроме материалов KeeleWeb, в Интернет помещено еще много «серьезных» словарей (термин Интернета, отделяющих их от развлекательных словарей, которых там хватает с лихвой). Галерею языков обогащают голландский, венгерский и русский. Eesti-hollandi sõnaraamat (Erkki Pilving; *www.ell.ee/~erkki/education/estned.html*) на первой странице предлагает лишь выбрать букву по алфавиту, чтобы затем читать соответствующую часть словаря (совсем как страницу книги). Количество слов невелико — 2960 слов. Базой для словаря послужили две книги (Hoogendijk, Van Nes, Tamm 1994; Prosa 1997). В том же ареале, но вне этого словаря, работала Анне Тамм, создавшая голландско-эстонскую базу данных, которую можно использовать для словаря в обоих направлениях (Tamm 1996). Ungari-eesti verbirektsioone (*www.ut.ee/Ural/UERS/*) представляет собой компьютерную версию напечатанного типографским способом словаря (Kippasto, Nurk, Seilenthal 1997), в которой можно по алфавиту направиться к нужной букве, вторая возможность — войти в словарную статью, воспользовавшись индексом на эстонском языке, который содержит все соответствия этого словаря. Eesti-vene-eesti sõnaraamat (ASE Computers; *www.ase.ee/dict/*)[4] объявляет своим объемом свыше 50 000 слов,

[3] Основа базы данных: Nigol 1998.

а приоритетным направлением — перевод с русского языка на эстонский: пользователь прежде всего видит отдельно вынесенные соответствия и уже затем словарную статью целиком. Технологический охват языка опять же невелик: поиск ограничен заглавными словами и все хоть немного завуалированные комбинации употребления (выражения, сочетания, части сложного слова и т.п.) в словарной статье остаются не найденными. Обратный перевод (эстонско-русская часть) предельно лаконичен, дается только соответствие.

В Интернете есть и другие справочные издания. Составители словарей все уютнее чувствуют себя среди новых методов теперешнего времени, которые меняют до неузнаваемости традиционную картотеку. Представители многих специальностей сами позаботились о доступности и наглядности своих данных и ввели в Интернет хорошо подготовленный материал. Проворнее других оказались, кажется, естествоиспытатели, они выступили с несколькими весьма ценными экспозициями. Например, Klassikalise geneetika leksikon (Mart Viikma; *www.cs.ut.ee/~martv*), в котором заглавное слово приводится вместе с английским соответствием и хорошим пояснением и по тексту которого с помощью гиперотсылок можно двигаться туда и обратно. Eestikeelsete taimenimede andmebaas (*www.ut.ee/BGBA/taimenimed. html*) предлагает примерно 8500 названий растений, одобренных соответствующей комиссией. Войти в эту базу данных можно путем поиска или через алфавит. Нельзя ошибиться в языке, т.е. забыть отметить, эстонский или латинский язык. Кое-что сделано и специально для школ, но здесь речь об этом не идет.

В некоторой мере представлена юридическая терминология: Ameerika õigusterminoloogia valiksõnastik (*www.ut.ee/jur_am_termin*), которым, однако, очень неудобно пользоваться из-за отсутствия единого механизма поиска. Этот материал отобран в одном специальном университетском курсе и потому относительно ограничен. То же касается материала еще одного учебника — Õiguse entsüklopeedia (Paul Narits; *www.ibs.ee/juura/entsyklop*).

Поскольку цель данного обзора — поподробнее представить словари, на этот раз вне внимания остаются другие интересные страницы: Eesti keele käsiraamat (*www.eki.ee/books/ekkr/*) с системой поиска; все богатые экспозиции эстонского фольклора (haldjas.folklore.ee), в том числе три базы данных — народная астрономия, народный юмор и поговорки (см. еще *www.eki.ee* и *www.ut.ee*).

Механизм поиска фольклора наглядно показывает, насколько важно в случае эстонского языка к системе поиска сразу привязать автоматическую морфологическую опору: по возможности гибко выстроенный запрос позволит среди прочего отметить и альтернативные буквы (*nai*[*ns*]*e*), чтобы выловить слова с изменяющейся основой (если исследователя привлекает на самом деле слово *naine*). То же относится и к газетам, журналам, вообще ко всем системам, в которых данные приходится извлекать из текстового массива: без морфологии поиск весьма беспомощен, а порой и вводит в заблуждение. Великий образец, направляющий компьютерное дело, английский язык не может здесь служить мерилом, потому что никак не подходит для таких богатых формами языков, как эстонский. С другой стороны, масса слов, перелопаченная беспомощным запросом, представляет собой лишь кучу хлама. Лучше справляются с задачей те системы, в которых вводимое слово понимается как сочетание-символ, а не как цельное слово. Такие проблемы возникают в связи с газетными архивами, которые для лексикографии, например при составлении толкового словаря эстонского языка, стали очень существенным дополнением имеющейся и неизбежно устаревающей картотеке литературного языка (более четырех миллионов карточек). Через Интернет составители-редакторы словарей попадают прямо в самую гущу современной периодики,

[4] Vene-eesti sõnaraamat I—IV, Tallinn 1984—1994. Авторские права на словарь принадлежат Институту эстонского языка. В компьютер (Интернет) введен неадекватный, с ошибками вариант.

получая возможность (как и в картотеке) исследовать слово в его контексте. Особенно важно это в случае новых слов, совсем отсутствующих в картотеке, но присутствующих как в активном употреблении, так и в других вариантах, т.е. исключение их из словаря было бы неверным шагом.

Хочу обратить внимание, что на странице словарей NETI (*www.neti.ee/cgi-bin/TEADUS/Filoloogia/Sonastikud/*) выделены четыре словаря с возможностью срочного запроса, так что можно вызвать слово уже там. Эти четыре словаря (на сегодня): эстонско-английский, эстонско-русский, английско-эстонский и словарь сленга. Пятая возможность — вызвать KeeleWeb, что расширит первичный запрос сразу на несколько словарей, перейдя непосредственно к уже описанному Ühispäring.

ЛИТЕРАТУРА

Goldberg, A. 1992, Argument Structure Constructions. PhD Dissertation (University of California), Berkeley.

—— 1995, Constructions. A Construction Grammar Approach to Argument Structure, Chicago.

Hanson, V., Tavast, A. 1996, Arvutikasutaja sõnastik. Inglise-eesti, Tallinn.

Hoogendijk, P., Van Nes, F., Tamm, A. 1994, Woordenlijst Nederlands-Estisch, Groningen.

Kippasto, A., Nurk, A., Seilenthal, T. 1997, Ungari-eesti rektsioonisõnastik, Tartu.

Loog, M. 1991, Esimene eesti slängisõnaraamat, Tallinn.

Nigol, S. 1998, Hargla murraku konsonantism, Tallinn.

Prosa, K. 1997, Eesti-hollandi vestmik. Estisch-nederlandse taalgids, Tallinn.

Pustejovsky, J. 1995, The Generative Lexicon, The MIT Press.

Saareste, A. 1979, Eesti keele mõistelise sõnaraamatu indeks, Uppsala.

Tamm, A. 1996, Dutsch and Estonian Bilingual Lexicography and Reversilbe Database Building of these Languages. Magistritöö, Tartu.

Vene-eesti sõnaraamat I—IV, Tallinn 1984—1994.

Viks, Ü. 1981, Eesti keele automaatsõnastikest. — KK, 211—216.

— — 1992, Väike vormisõnastik I. Sissejuhatus & grammatika; II Sõnastik & lisad, Tallinn.

Väike murdesõnastik I—II, Tallinn 1982—1988.

Õim, A. 1991, Sünonüümisõnastik, Tallinn.

—— 1993, Fraseoloogiasõnaraamat. Tallinn.

—— 1995, Antonüümisõnastik. Tallinn.