СТРУКТУРА И СОДЕРЖАНИЕ ЛЕКСИЧЕСКИХ БАЗ ДАННЫХ ДЛЯ ОБУЧАЮЩЕЙ ПРОГРАММЫ ПО ИНОСТРАННЫМ ЯЗЫКАМ

STRUCTURE AND CONTENT OF LEXICAL DATABASES FOR A FOREIGN LANGUAGE TEACHING SOFTWARE

М.Н. Михайлов *(mihail.mihailov@uta.fi)*
Тамперский университет, Финляндия

Существующие обучающие программы по иностранным языкам в основном представляют собой компьютерную реализацию традиционных «бумажных» сборников упражнений. В настоящей статье показано, что хорошо структурированная словарная база данных значительно повышает потенциал такого рода учебных материалов.

Компьютер в учебном процессе

Последнее десятилетие ознаменовалось активным включением вычислительной техники в учебный процесс. Создаются обучающие программы по различным дисциплинам, проводится тестирование с использованием вычислительной техники, разрабатываются пакеты программ для дистанционного обучения (см. Holmberg et al. 2005). Темпы и формы развития в этой сфере в первую очередь зависят от доступности вычислительной техники и Интернета в учебных заведениях и в быту.

Создание обучающих программ по иностранным языкам представляется одной из самых важных сфер компьютеризации обучения. Изучение языков связано с большими затратами времени и усилий, при этом существует необходимость индивидуальной и самостоятельной работы. Внедрение обучающих программ в учебный процесс позволило бы несколько разгрузить преподавателя и повысить результативность обучения (Higgins 1988; Higgins, Johns 1984).

Обучающие программы по иностранным языкам: state of the art

Оценивать положение дел в этой сфере крайне трудно, поскольку существует довольно большое количество коммерческих пакетов программ, доступ к которым ограничен. Поэтому во многих случаях мне приходилось довольствоваться описаниями программных продуктов или демонстрационными версиями. Однако, представляется, что в целом существующие тенденции проследить удается.

Все обучающие программы по иностранным языкам можно разделить на три категории:

Пустые оболочки для создания рабочих глоссариев / упражнений

Программы-тесты

Обучающие игры

Оболочки для глоссариев, по-видимому, обычно создаются программистами, интересующимися изучением иностранных языков и не являющимися профессиональными лингвистами, поэтому лингвистическая компонента часто оставляет желать лучшего. Как правило предусмотрен лишь ввод пар «слово – эквивалент», иногда можно добавить картинку, но ни в одной из известных мне программ нельзя давать примеры употребления, не говоря уже о толкованиях и грамматической информации (см. напр. Hot Potatoes, KVocTrain, Open Book). Наиболее интересным решением представляется включение утилиты для заучивания слов в пакет программ, обслуживающих электронные словари, как это сделано, например, в словарях Lingvo, для которых разработан лексический тренажер Lingvo Tutor (см. ABBYY Lingvo).

Программы-тесты создаются преподавателями иностранных языков в сотрудничестве с программистами. Большинство таких программ представляют собой электронные реализации «бумажных» упражнений по лексике и грамматике. Упражнения и тесты строятся по шаблонам типа «Вставьте слово в правильной форме» или «Выберите правильный ответ». Упражнения могут быть разбиты по темам или по уровню сложности. Недостаток у них тот же, что и у традиционных сборников упражнений: они конечны и студент лишь в редких случаях может получить комментарии по своим ошибкам. Вызывает сомнение и рентабельность такого «автомата для переворачивания страниц» («page-turner», Mohan 1991: 112).

Игровые программы представляют собой компьютерные реализации языковых игр: кроссвордов, Scrabble, «Виселицы» и т.п. — и сюжетные игры (см., напр., Higgins 1988; Higgins, Johns 1984). Такие программы позволяют развивать некоторые языковые умения и расширяют словарный запас, но при этом мало помогают в обучении грамматике (Cheung, Harrison 1992).

Таким образом, главные недостатки существующих обучающих программ — «конечность» и «шаблонность» заданий. Это связано главным образом с тем, что доля лингвистики в этих продуктах незначительна. Поэтому объем данных или оказывается недостаточно большим или выходит из-под контроля. В

результате на текущий момент роль обучающих программ в обучении иностранным языкам лишь вспомогательная.

SysMLL – обучающая система нового поколения

В Тамперском университете (Финляндия) силами нескольких подразделений (Институт современных языков и переводоведения, Центр обучения иностранным языкам, Отделение вычислительной техники, Виртуальный университет) создается обучающая система по иностранным языкам SysMLL (Multilingual System for Language Learning). Цель системы — расширение, активизация и тестирование словарного запаса студентов.

Сначала будет создан русско-финский модуль, затем будет добавлен английский, позднее возможно пополнение системы и другими языками[1]. Доступ к системе будет осуществляться через веб-интерфейс.

Главное отличие этой системы от существующих состоит в том, что в основе ее не шаблон для ввода упражнений/тестов, а большой структурированный словарный массив, используя который система сама генерирует тесты и упражнения. Это позволяет сделать работу с тренажером «бесконечной». Задача преподавателя сводится лишь к вводу лексики для активизации. Пользователи системы также получают возможность вести свои рабочие глоссарии, часть этого материала может позднее включаться администратором базы данных в основной словарь.

В настоящей публикации я не буду рассматривать технические вопросы, связанные с разработкой интерфейса и алгоритмами, а остановлюсь на лингвистических аспектах, а именно – на создании словарного массива для обучающей программы.

Традиционный словарь — за основу?

Довольно заманчивой может показаться следующая идея. Берем хороший толковый и/или двуязычный словарь, загружаем его в базу данных — и система «заряжена». Такой подход в частности применяется для заполнения готовых языковых модулей в лексическом тренажере Open Book (http://www.vinidiktov.ru/openbook.htm). Несмотря на то, что таким образом можно быстро получить большие массивы данных (хотя при этом, возможно, возникнут проблемы, связанные с авторскими правами), использование готовых словарей представляется проблематичным. Дело в том, что словари (даже т.н. учебные словари) изначально не предназначены для их выучивания наизусть, а являются лишь справочным материалом.

Возьмем в качестве иллюстрации статью из англо-русского словаря (Lingvo Universal, http://www.lingvo.ru/lingvo/):

Firefly [транскрипция]
сущ.
светляк (летающий)
Syn:
lightning bug (1)

В статье содержится информация, достаточная для того, чтобы пользователь словаря смог бы понять, что оно значит, как оно произносится, какие у него синонимы. Однако для обучающей системы информации оказывается недостаточно. Во-первых, отсутствует другой, более употребительный эквивалент *светлячок* (слово *светляк* встречается в литературе 19-го века, а также в составе научных названий, например, *светляк большой*). Во-вторых, нет толкования, картинки, примеров употребления.

Продолжим обсуждение той же лексемы и посмотрим статью из толкового словаря. В словаре С.И. Ожегова и Н.Ю. Шведовой читаем следующее:

СВЕТЛЯК, *а* и **СВЕТЛЯЧОК**, *чка*, *м*. Жучок (а также его личинки и яйца), светящийся в темноте. (2)

Здесь имеется грамматическая информация, есть толкование, но нет примеров употребления. Кроме того, у статьи два лексических входа, то есть по-видимому предполагается, что *светляк* и *светлячок* — полные синонимы, что на мой взгляд не совсем верно (см. выше).

Большие проблемы могут возникнуть с «отсылочными» словарными статьями типа:

ПОКУПАТЬ[2] см. купить. (Ожегов, Шведова 2003) (3)

Наконец, не совсем понятно, что делать с рядами приблизительных эквивалентов. Например, Lingvo Universal предлагает для английского прилагательного *weird* в одном из его значений следующую цепочку: *странный, жуткий, непонятный; причудливый, фантастический*. Если этот ряд попадет в обучающую систему в своем изначальном виде, то пользователь должен будет основательно тренировать свою память, чтобы запомнить все эти прилагательные именно в этой последовательности.

Таким образом, база данных для эффективно работающего лексического тренажера должна довольно сильно отличаться от традиционных словарей по своему составу и организации.

Вопрос о единице хранения данных

Как известно, основной единицей хранения для традиционного словаря, как правило, является лексема. Но при создании многоязычных лексических баз данных такой подход оказывается очень неудобным. Поэтому в компьютерной лексикографии развиваются новые способы представления лексикографических данных,

[1] Данная статья для удобства восприятия будет в основном иллюстрироваться русскими и англо-русскими примерами.

позволяющие решить вопрос о семантических и сочетаемостных различиях межъязыковых соответствий, а также проблему лакун (Boas 2005, Janssen 2004, Martin 2004, Fellbaum 1998).

При обучении иностранным языкам основной единицей также является не лексема, а скорее лексико-семантический вариант (ЛСВ), то есть комплекс «слово + значение». Например, вышеупомянутое английское слово *weird* в своем разговорном значении 'странный, непонятный' может быть вполне полезно уже на среднем этапе обучения для понимания фраз типа *He's got some weird ideas*. Однако устаревшее значение этого слова 'роковой, фатальный' (например, шекспировские *weird sisters* из «Макбета») вряд ли следует изучать раньше продвинутого этапа.

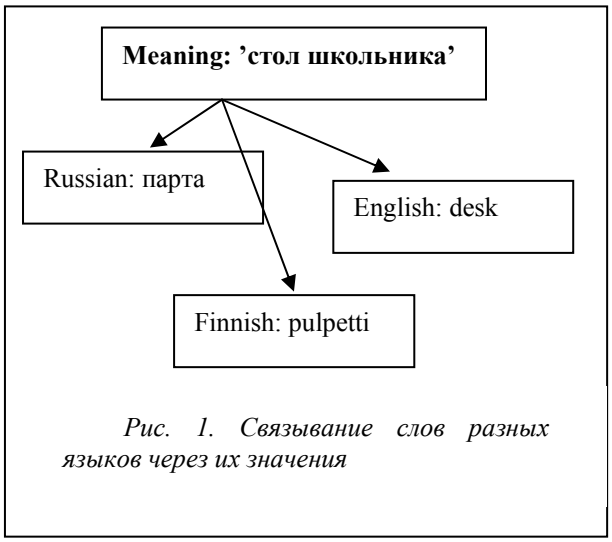

Рис. 1. Связывание слов разных языков через их значения

Разные значения одного и того же слова могут попадать в разные тематические группы, например слово *рыба* в одном значении относится к группе «Фауна», в другом — к группе «На работе» (*подготовить рыбу для доклада*), в третьем — к группе «Игры» (положение в домино). Кроме того, разные значения одного и того же слова могут различаться по уровню сложности, например, слово *рыба* в первом значении несомненно входит в базовый лексикон, остальные упомянутые выше значения уже относятся к продвинутому уровню. У разных значений одной и той же лексемы могут быть разные соответствия в других языках, у некоторых из значений может вообще не оказаться точных эквивалентов (ср. рус. *рыба* — англ. *fish*, *template*[2]*, block*).

По этим причинам многозначные лексемы представлены в базе данных SysMLL в виде нескольких записей, причем наиболее удобным для многоязычной системы оказывается подход от значения к форме. Семантика оказывается как бы «мостиком», соединяющим разные языки (см. рис. 1). Таким образом, в системе используется нечто вроде interlingua, но задачи построения четкого метаязыка для описания семантики слова пока не ставится. При построении базы данных для лексического тренажера это оказывается вполне возможным, поскольку отсутствует необходимость организации полномасштабного поиска (как, например, в электронном словаре).

Такой способ представления данных оказывается гибким и экономным. Появляется возможность работы с разными языковыми парами. Данные по лексике каждого из языков можно описывать отдельно, многозначность описываемых лексем также серьезных проблем не создает. Так, в примере на рис. 1 многозначное английское слово *desk* задействовано только в значении 'школьная парта'.

Появляется также и возможность решить проблему безэквивалентной лексики: для таких слов «мостик» к другим языкам будет отсутствовать, и работа с ними будет вестись только в одноязычном режиме (например, от толкования или графического / звукового образа).

Содержание словарной статьи

Традиционные словари всегда ограничены по объему: с очень толстым или многотомным словарем неудобно работать, очень длинную словарную статью трудно читать. Для электронных словарей ограничений значительно меньше, а в базе данных обучающей программы может быть заложено даже еще больше информации: ведь не все будет предъявлено пользователю, это лишь материал для порождения упражнения.

Отбор лексики

Обучающий словарь не предполагается использовать в качестве справочной базы данных. Поэтому требование полноты для продуктов такого рода сводится к полноте представленности лексики по отдельным темам и уровням сложности. Это дает возможность развивать систему поэтапно, добавляя в нее все новые тематические группы или пополняя и уточняя уже имеющиеся. При отборе лексики можно в целом руководствоваться теми же принципами, что и при составлении словарей, например, используя данные по частотности лексики в текстах искомой категории сложности. Тем не менее, слова некоторых частей речи в обучающий словарь не попадут. Так, нет особого смысла включать в словник служебные слова: предлоги, союзы, частицы. Знакомство с числительными и местоимениями также входит в обязательный минимум, который должен быть усвоен до начала работы с программой.

Грамматика

Очень важно, чтобы студент с самого начала имел правильную информацию о грамматических признаках слова, например:

грамматический род существительного, особенно в неявных случаях (рус. *дверь* — ж.р., *зверь* — м.р.),

основные формы неправильного глагола (англ. *fight — fought — fought*),

нерегулярные формы (рус. *брат — братья*),

парные лексемы, например, русские видовые пары (рус. *покупать — купить*, *брать — взять*), или обозначения лиц мужского и женского пола (рус. *спортсмен — спортсменка*)

и т.д. и т.п.

Поскольку возможность создавать отсылочные словарные статьи отсутствует, вся необходимая информация должна быть представлена в виде одной единицы хранения[2].

Рубрикация и уровни

Разучивание слов в случайном порядке малоэффективно. Новые слова должны встраиваться в уже имеющиеся ассоциативные ряды, обрастать иллюстративными примерами, зрительными образами. Поэтому для каждого слова, зафиксированного в базе данных, должна быть указана тематическая рубрика. Набор таких рубрик является открытым множеством, структура базы данных позволяет модифицировать рубрикацию.

Другой важный момент — уровень сложности. Применение этого параметра дает возможность использования системы пользователями, в разной степени владеющими изучаемым языком. Кроме того, появляется возможность проведения тестирования словарного запаса.

Толкования, картинки и примеры

Тренировка на знание переводных эквивалентов может быть очень полезна для начального уровня, а также при подготовке переводчиков. Однако наряду с разучиванием двуязычных соответствий важно развивать и навыки объяснения слов изучаемого языка. Интересно, что словарные дефиниции далеко не всегда оказываются пригодными для учебных целей, например, толкование слова *жаворонок* «Певчая птичка отряда воробьиных» (Ожегов, Шведова 2003) недостаточно иллюстративно, поэтому нередко возникает необходимость уточнения / упрощения / дополнения толкований.

Использование мультимедийных средств — изображения и звука — часто дает возможность объяснять значение слов, плохо поддающихся толкованию. Конечно, не все слова удается достаточно наглядно проиллюстрировать. Так, упомянутый выше *жаворонок* визуально распознается гораздо хуже *лебедя* или *страуса*.

Слова лучше запоминаются в контексте, поэтому очень желательно снабжать слова иллюстративными примерами употребления. Примеры должны быть короткими, понятными и хорошо запоминающимися[3].

Программа-помощник

Если в словаре системы содержится такая богатая и разнообразная информация о словах, создание упражнений и тестов — лишь дело техники.

Легко реализуется режим предварительного ознакомления с лексикой: просмотр списков слов, переводных эквивалентов и т.д.

В режиме тренировки программа может предложить пользователю указать переводной эквивалент для слова (предъявляя, кроме самого слова, картинки или пояснения, чтобы не возникло проблем с многозначностью и омонимией), подобрать слово или выражение, соответствующее данному толкованию или картинке, вставить пропущенное слово в правильной форме и т.п.

Кроме того, вполне реализуемыми оказываются и различные языковые игры — кроссворды, чайнворды и т.п.: слова выбираются из базы данных, а ключами служат толкования / картинки / переводные эквиваленты.

Система будет работать в многопользовательском режиме. По каждому пользователю программа будет собирать информацию о словах, вызывающих трудности. Эти слова будут предъявляться пользователю чаще, чем другие, до тех пор, пока тот не перестанет ошибаться.

По результатам работы студента система будет способна генерировать отчет для преподавателя или самого студента:

количество сеансов работы, общее время работы,
результативность, процент ошибок,
прогресс / регресс,
слова / темы, вызывающие затруднения,
типичные ошибки,
рекомендации,
общая оценка работы.

Заключение

Таким образом, словарная база данных для обучающей программы по лексике оказывается лексикографическим продуктом нового типа. Использование традиционных толковых и двуязычных словарей в качестве «готовых» баз данных мало что дает, поскольку они организованы для других целей и по другим принципам. Во всяком случае, при переносе данных потребуется так много изменений – и в составе словника, и в метаязыке, и в членении материала, – что в конечном итоге все равно получится совершенно новый продукт.

Создание языковых ресурсов такого типа приводит к появлению нового поколения обучающих программ. Наличие хорошо структурированных словарных баз данных, содержащих многообразную информацию о значении слов, особенностях словоизменения, употребления, о соответствиях в других языках, позволяет создавать обучающие программы, которые способны сами генерировать упражнения и тесты и контролировать их выпол-

[2] Разумеется техническая реализация этой единицы хранения в базе данных достаточно сложна, и информация по одному ЛСВ оказывается разбросана по нескольким таблицам.

[3] Поэтому пока неясно, насколько реально подключение к такой системе корпуса текстов для поиска примеров употребления.

нение. Хотя исходной точкой является создание лексических тренажеров, наличие в базах данных грамматической информации позволяет организовывать тренинг и по грамматике при наличии соответствующего программного обеспечения (см. тж. Михайлов, Сидоров 1995).

Использование вычислительной техники в преподавании иностранных языков несомненно повысит эффективность и результативность обучения. Тем не менее, обучающие программы, как бы хороши они ни были, все равно останутся только помощниками. Учителями иностранных языков они не станут.

Литература

1. *Михайлов М.Н., Сидоров Г.О*. Тренажер по русской грамматике. // Мультимедиа в преподавании языков. Тезисы конференции. М.: МГУ, 1995.
2. *Ожегов С. И., Шведова Н. Ю.* Толковый словарь русского языка. М.: Азбуковник, 2003.
3. *Boas H. C.* Semantic Frames as Interlingual Representation for Multilingual Lexical Databases // International Journal of Lexicography, 18(4), 2005. – Pp. 445–479.
4. *Cheung A., Harrison C.* Microcomputer Adventure Games and Second Language Acquisition: A Study of Hong Kong Tertiary Students // Pennington M. C., Stevens V. (eds.). Computers in Applied Linguistics: an International Perspective. Clevedon – Philadelphia – Adelaide: Multilingual Matters, 1992. – Pp. 155–181.
5. *Fellbaum C*. WordNet: an Electronic Lexical Database. Cambridge, Mass.: MIT Press, 1998.
6. *Higgins J.* Language, Learners, and Computers. London – New York: Longman, 1988.
7. *Higgins J.*, Johns T. Computers in Language Learning. Collins ELT, Addison-Wesley, 1984.
8. *Holmberg B., Shelley M., White C. (eds.)* Distance Education and Languages. Evolution and Change. Clevedon – Buffalo – Toronto: Multilingual Matters, 2005.
9. *Janssen M*. Multilingual Lexical Databases, Lexical Gaps, and SIMuLLDA // International Journal of Lexicography, 17(2), 2004. – Pp. 137–155.
10. *Martin W*. SIMuLLDA, the Hub-and-Spike Model and Frames or How to Make the Best of Three Worlds. // International Journal of Lexicography, 17(2), 2004. – Pp. 175–189.
11. *Meijis W*. Computers and Dictionaries // Butler C.S. (ed.) Computers and Written Texts. Oxford UK & Cambridge USA: Blackwell, 1992. – Pp. 141–167.
12. *Mohan B.* Models of the Role of the Computer in Second Language Development // Pennington M. C., Stevens V. (eds.). Computers in Applied Linguistics: an International Perspective. Clevedon – Philadelphia – Adelaide: Multilingual Matters, 1992, – Pp. 110–127.

Интернет-ссылки

1. ABBYY Lingvo: http://www.lingvo.ru/ multilingual/
2. Hot Potatoes: http://web.uvic.ca/hrd/halfbaked/
3. KVocTrain: http://edu.kde.org/kvoctrain/
4. Open Book: http://www.vinidiktov.ru/openbook.htm