Е.С. Чернобривец (Минск, БГУ)

ЛИНГВИСТИЧЕСКИЕ ОСОБЕННОСТИ КОРПУСА «ТАТОЭБА»

Проект «Татоэба» 例えば(«Например» — по-японски) — сайт для свободного обмена примерами фраз на всех доступных языках мира.

Создателем и лидером проекта является Чанг Хо, француженка вьетнамского происхождения. В своем профиле Чанг Хо — создательница проекта — указывает, что является носителем французского языка, бегло говорит по-английски, знает вьетнамский, так как это язык ее матери, а также учила испанский, японский, немецкий и китайский языки в университете и самостоятельно.

Позже к Чанг Хо присоединился второй администратор, Симон Аллан, также француз. Первые примеры сайта датированы 30 сентября 2007 г.

Стимулом к созданию данного проекта, по словам Чанг Хо, была любовь к иностранным языкам и неспособность найти хороший он-лайн словарь. Начальным ресурсом для сайта стал англо-японский корпус языковых примеров профессора Ясухиро Танаки. В декабре 2010 г. «Татоэба» содержал более 648 000 предложений.

На данный момент сайт содержит 1 млн 789 тыс 165 предложений на 114 языках мира. С каждым днем эти показатели растут.

Корпус «Татоэба» — это не обычный словарь, который переводит слова — это словарь, который переводит цельные семантические конструкции — предложения, фразы, пословицы, поговорки, и, следовательно, является словарем предложений. Эти фразы представлены в корпусе на различных языках.

Почему именно предложения? Потому что предложения интереснее; они имеют яркий контекст, они могут научить нас большему, чем просто слова.

Следующая особенность — это мультиязычность словаря «Татоэба». Как уже упоминалось, корпус содержит 114 языков. Самые популярные из них: английский, эсперанто, японский, французский, немецкий, испанский, турецкий, португальский, итальянский и др. Существует более тысячи предложений на следующих языках: литературный арабский, исландский, хинди, уйгурский, вьетнамский, норвежский (букмол), белорусский, шанхайский и кантонский

диалекты китайского языка. В проекте также фигурируют и языки искусственные: эсперанто, клингон, интерлингва, CycL, топикона.

Первоначально для введения нового языка было достаточно лишь обратиться к администраторам и внести на нём пять примеров. Впоследствии необходимым требованием стало наличие сертификации вносимого языка согласно стандарту ISO 639-3 (Набор стандартов Международной организации по стандартизации, связанный со стандартизацией названий языков и языковых групп).

При заявке на добавление нового языкового раздела можно предложить флаг-символ, который его обозначит на сайте; этот графический знак не обязан представлять конкретное государство из современных либо существовавших ранее.

В корпусе «Татоэба», в отличие от других электронных словарей, все переводные соответствия взаимосвязаны, и «лишней пары» возникнуть не может, что, несомненно, является большим преимуществом.

Значительной особенностью словаря «Татоэба» является его открытость и общедоступность. Все предложения из корпуса доступны для загрузки под свободной лицензией CC-BY (Creative Commons Attribution). Данная лицензия предполагает, что пользователь может распространять, изменять, а также использовать материал в коммерческих целях при обязательном условии — указывать источник и автора. Таким образом, «Татоэба» является некоммерческим проектом, и любой желающий вне зависимости от специализации и языковой принадлежности может вносить изменения в базы данных проекта (добавлять или редактировать существующие фразы).

Зарегистрироваться и дополнять примеры фраз на родном или изучаемом языке может любой желающий с базовыми навыками грамотности. Просматривать накопленный материал могут все желающие, добавлять и редактировать — только зарегистрированные участники. Участники со стажем могут получить статус «trusted user» («доверенный»). Он дает доступ к тегам, в также позволяет связывать адекватные переводы между собой или «исключать» неадекватные. Ограниченный круг участников проекта имеет статус «corpus maintainers» («блюстители»), имеющих администраторские полномочия.

Принцип работы проекта заключается сборе и увязывании переводов той или иной фразы данного языка. Это могут быть, например, разговорные фразы, вопросы, пословицы и поговорки, любые другие связные предложения. Система анализирует все поступившие данные. Если конструкция A переводится на другой язык как конструкция B, а та, в свою очередь — как конструкция C, то все они будут считаться переводами друг друга, обозначающими одно и то же. После этого при запросе одной из них отображены будут все три.

Корпус «Татоэба» имеет структуру графы, как показано на рис. 1. Каждое предложение имеет связь с предложением на другом языке, и если эта связь двусторонняя, то эти предложения имеют одинаковое значение.

Для собственных переводов с помощью «Татоэба» рекомендуется ориентироваться только на оригинал, так как сопутствующие переводы могут быть

неточны. Обсуждение нюансов перевода возможно в комментариях к каждому из предложений.

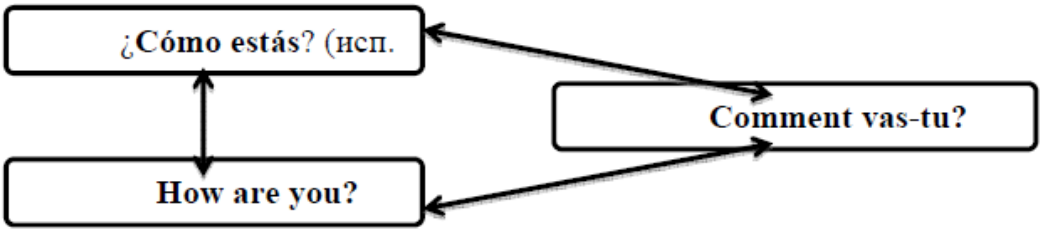

Рис. 1. **Структура корпуса «Татоэба» на примере фразы «Как дела?»**

Для определения языка используется программа «Tatodetect«. Основой это программы являются n-граммные модели языка. Создателем этой программы является Алан Симон. Словарь содержит частые n-граммы для всех рассматриваемых языков. Вначале программа пробует определить язык по пентаграммам, т.е. по сочетаниям пяти символов. Если в словаре недостаточно пентаграмм, производится проверка с использованием тетраграмм, триграмм и биграмм. Это позволяет эффективно работать как с европейскими языками, так и с восточноазиатскими. Транскрипция для кантонского китайского и путунхуа, а также шанхайского языка генерируется с помощью созданных Симоном Алланом специальных приложений.

Критика корпуса «Татоэба» связана с несколькими неоднозначными моментами в его устройстве:

1) Языки на сайте обозначаются флагами: иногда флагом государства, где говорят на языке, иногда выдуманным флагом. Некоторые флаги спорны (например, жители США высказывались против обозначения английского флагом Великобритании).

2) Среди примеров есть достаточно большое количество одинаковых или незначительно отличающихся предложений.

Как признает Чанг Хо, проект «Татоэба» несовершенен и нуждается в постоянных корректировках и доработках. Поэтому администрация сайта призывает всех пользователей принимать активное участие в исправлении ошибок и неточностей, а также приветствует обратную связь.

Правила пользования проектом «Татоэба» перечислены в статье «*How to be a good contributor in Tatoeba*», размещенной в блоге корпуса. С этими правилами должен ознакомиться каждый, кто желает участвовать в проекте. Некоторые из них перечислены ниже:

- не переводить предложения дословно;
- не менять языки предложений — это нарушит структуру переводов;
- принятие правил лицензии CC-BY;
- сообщать администрации корпуса об ошибках и др.

Соблюдение этих правил и советов обеспечивает организованность проекта, а также позволяет минимизировать количество ошибок и неточностей в переводах.

Проект «Татоэба» выступает за открытость и свободу Интернета в проекте «Mozilla Drumbeat» вместе с другими проектами-участниками. Он также со-

действует многим электронным словарям и переводчикам –например, электронному словарю японского языка «WWWJDIC». Проект «Татоэба» сотрудничает с проектом «Shtooka» — бесплатной коллекцией аудиозаписей слов, фраз, пословиц и т.п. на различных языках.

Благодаря своим особенностям проект «Татоэба» постепенно получает признание уникального средства для самостоятельного обучения, цель которого — сделать Интернет более удобным местом для изучения языков.

ЛИТЕРАТУРА

1. Корпус Татоэба [Электронный ресурс]. — Режим доступа: http://tatoeba.org. — Дата доступа: 18.02.2013.

2. Википедия, свободная энциклопедия [Электронный ресурс]. — Режим доступа: ru.wikipedia.org. — Дата доступа: 20.02.2013.

3. Блог Татоэба, How to be a good contributor in Tatoeba [Электронный ресурс]. — Режим доступа: http://blog.tatoeba.org/2010/02/how-to-be-good-contributor-in-tatoeba.html#rule1. — Дата доступа: 21.02.2013.

4. Introduction to the Tatoeba.org Project [Электронный ресурс]. — Режим доступа: http://www.youtube.com/watch?v=ac9SmJuwHqk. — Дата доступа: 23.02.2013.