

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Филологический факультет

АКТУАЛЬНЫЕ ПРОБЛЕМЫ ЛИНГВИСТИКИ И ЛИТЕРАТУРОВЕДЕНИЯ

Сборник материалов IX (XXIII)
Международной научно-практической
конференции молодых учёных
(14–16 апреля 2022 г.)

Выпуск 23

Томск
Издательство Томского государственного университета
2022

doi: 10.17223/978-5-907572-04-1-2022-46

ПРИНЦИПЫ РАЗМЕТКИ ОШИБОК ПРИ СОЗДАНИИ УЧЕБНОГО КОРПУСА РУССКИХ ТЕКСТОВ

Хоу Л.

Российский государственный педагогический университет им. А.И. Герцена,
аспирант

PRINCIPLES OF ERROR TAGGING FOR BUILDING RUSSIAN LEARNER CORPORA

Hou L.

Herzen State Pedagogical University of Russia, postgraduate student

В статье рассматривается актуальная проблема разметки ошибок, которая является трудоемкой задачей, выполняемой при создании учебных корпусов русских текстов. Обобщаются основные принципы разметки, принятые при создании учебного корпуса. Основное внимание уделено выяснению установлению принципов разметки ошибок в русских текстах в сочетании с характеристиками учебного корпуса и знаний китайских студентов.

Ключевые слова: учебный корпус, разметка, принципы, ошибки

The paper concerns the issues of annotation when building Russian Learner corpora. The basic principles of annotation are summarized. The emphasis is on the elucidation and description of standards for error tagging based on the characteristics of Learner corpora and knowledge of Chinese students.

Key words: learner corpus, tagging, principles, errors

Научный руководитель: Л.Н. Беляева, д-р филол. наук, профессор РГПУ им. А.И. Герцена.

Решение задачи усовершенствования процесса обучения иностранному языку основывается сегодня на применении различных инструментов информационных технологий. Одним из таких инструментов является учебный корпус, внедрение которого в педагогическую деятельность начинает привлекать внимание преподавателей [1. С. 98]. К ним относятся китайские ученые, чьи исследования по учебным корпусам в основном сосредоточены на англоязычных корпусах. Однако, по сравнению с корпусами английского языка, разработка учебного корпуса русского языка в Китае идет относительно медленно, и отсутствуют соответствующие

стандарты построения корпуса, принципы разметки ошибок, опыт его применения и т.д. Цель данной статьи – рассмотрение принципов разметки ошибок при создании китайского исследовательского учебного корпуса русских текстов. Данный корпус создается усилиями преподавателей и студентов кафедры русского языка Второго Пекинского университета иностранных языков. В корпусе включаются 164 сочинения 41 китайских студентов 1-го курса по специальности «Русский язык». Общий объем корпуса составляет 26 400 токенов – слогоморфем.

Под учебным корпусом поднимается электронный сборник аутентичных текстов, написанных лицами, изучающими иностранный язык или второй язык [2. Р. 12]. Аутентичность языковых материалов, большой объем материала обучающихся и удобные средства поиска на основе корпусных и компьютерных технологий – все это делает учебный корпус важным ресурсом для изучения и преподавания иностранного языка. Разметка является ключевой частью отражения разнообразных данных о представленных в корпусе материалах [3. С. 45] и придает корпусу особую ценность. Под разметкой (англ. *tagging, annotation*) понимается процесс или результат приписывания текстам и их компонентам специальных тегов [4. Р. 310]. Разметка даёт возможность идентифицировать тексты по различным параметрам, позволяя осуществлять осмысленный поиск по корпусу [5. С. 514].

Разметка ошибок (англ. *error tagging*), которая выявляет «слабое место» учащегося, является не только уникальным типом разметки, характерным именно для учебного корпуса, но и одним из наиболее важных типов разметки корпуса [6. Р. 18]. На основе разметки ошибок посредством инструментов поиска можно значительно повысить эффективность извлечения и статистики ошибок, всесторонне и систематически отразить распределение ошибок учащихся. Можно сказать, что разметка ошибок актуализирует применение учебного корпуса на практике.

Основные принципы разметки корпуса были предложены Джеффри Личем [7. Р. 277]:

- разметка должна быть независима от текста: должна быть возможность убрать разметку и просмотреть текст без неё;

- принципы разметки и их разработчики должны быть известны конечному пользователю;
- пользователь должен быть поставлен в известность о том, что разметка представляет собой лишь потенциально полезный инструмент;
- в основу разметки должны быть положены общепринятые и, по возможности, теоретически нейтральные лингвистические принципы;
- ни одна разметка не может априорно считаться стандартом.

Резюмируя, следует отметить, что разметка тесно связана с применением корпуса, она должна быть максимально удобной для пользователей. Следовательно, для облегчения практического применения теги разметки ошибок должны отличаться от представления текстов в корпусе. При разметке в учебном корпусе русских текстов ошибки помечаются строчными латинскими буквами и цифрами, включенными в угловые скобки <...>, например: *Я уже два года не виделась <vp1> друзей.* Помимо этого, при этом подходе разметка ошибок не изменяет исходный корпус, что позволяет в наибольшей степени сохранить аутентичность корпуса.

Кроме вышеуказанных принципов разметка ошибок учебного корпуса должна учитывать такие аспекты, как классификация ошибок, метод разметки, характеристика ошибок учащихся и т.д. На основе изучения знаний китайских студентов по специальности «Русский язык» и с учетом перечисленных выше аспектов предлагаются следующие принципы разметки ошибок в учебном корпусе русских текстов.

1. Простота тегов разметки обеспечивает их практическую применимость.

Механизм разметки корпуса может быть автоматическим, ручным или автоматическим с ручной корректировкой [8. С. 48]. Для учебного корпуса трудоемкой задачей является процесс обнаружения и разметки ошибок в тексте, который обычно проводится вручную. Ручная разметка требует много времени и силы, ввиду этого теги разметки должны быть краткими и простыми в применении. С точки зрения пользователя, чем более

подробно аннотирован корпус, тем лучше, но для практики также необходимо учитывать выполнимость и целесообразность разметки. Детальность разметки ошибок способствует информативности извлечения ошибок, но стремление к чрезмерно детальной разметке приводит к появлению слишком большого количества тегов. Это не только усложняет ручную разметку, но и не способствует ее практическому использованию.

2. Иерархическое устройство разметки обеспечивает обозримость классов ошибок.

Выбору модели разметки (собственно, выбору списка тегов) предшествует решение задачи квалификации ошибки [9. С. 176]. На основе анализа материалов учебного корпуса была разработана базовая классификация ошибок, ошибки разделены на 4 основных типа и 14 подкатегорий. В соответствии с методом классификации ошибок устройство разметки ошибок должно быть иерархическим, обеспечивающим обозримость классов ошибок учащихся [10. С. 150]: например, орфографические ошибки отмечаются строчными латинскими буквами <fm>, неверное словообразование и неверное использование прописной буквы, которые относятся к тому же типу ошибок, помечаются тегами <fm1>, <fm2>.

3. Возможность индивидуально настраивать разметку на основе практической цели исследования.

Как правило, разметка учебного корпуса связана с особенностями его применения, хотя, естественно, можно выделить типы разметки (метаразметка, морфологическая разметка и др.), которые являются универсальными. В случае учебных корпусов в зависимости от планируемого применения корпуса уровень и направленность разметки различны и индивидуальны [11. Р. 2]. В случае учебного корпуса русского языка китайских студентов следует учитывать, что большинство студентов, обучающихся по специальности «Русский язык» в китайских университетах, изучало английский как второй иностранный язык в начальной, средней и старшей школе. Поэтому на первом этапе изучения русского языка им часто мешает английский язык, так как возникает интерференция. Некоторые студенты изучают русский язык со

старшей школы, поэтому английский язык на них влияет меньше. Таким образом, следует учитывать различие второго иностранного языка учащихся и степень владения языком при разметке для анализа причины возникновения ошибок.

Рассмотренные выше принципы разметки ошибок реализованы при решении задач создания исследовательского учебного корпуса русских текстов. Этот корпус представляет базу для дальнейшего анализа ошибок, допускаемых китайскими студентами по специальности «Русский язык».

В целом можно утверждать, что разметка ошибок является обязательной и ключевой частью при создании учебного корпуса, а также важным предметом исследования в этой области. Используя данную разметку, преподаватели могут быстро и точно анализировать ошибки, допускаемые учащимися в процессе овладения иностранным языком, чтобы постоянно корректировать свою педагогическую деятельность для повышения эффективности результатов учебного процесса [12. С. 65]. Кроме того, разметка ошибок в учебном корпусе предоставляет самим учащимся богатые примеры и реальные контексты ошибок, чтобы они могли легко принимать и запоминать языковые правила иностранного языка.

Список источников

1. Ольховская А.И. Корпусное преподавание русского языка // Вестник Томского государственного педагогического университета. 2019. Вып. 2 (199). С. 98–107.
2. Grigaliūnienė J. Corpora in the classroom. Vilnius University, 2013. 81 p.
3. Замичева С.С., Иванцова Е.В. Тематическая разметка диалектного корпуса: опыт Томских диалектологов // Вестник Томского государственного университета. Филология. 2020. № 67. С. 45–61.
4. Foster M.E., Oberlander J. Corpus-based generation of head and eyebrow motion for an embodied conversational agent // Language Resources and Evaluation. 2007. № 41. P. 305–107.
5. Елохов Е.С., Югов А.С., Ланин В.В. Создание хранилища научных публикаций с применением методов корпусной лингвистики // Открытые семантические технологии проектирования интеллектуальных систем. 2014. № 4. P. 513–516.

6. Granger S. A bird's-eye view of learner corpus research // Computer Learner corpora Second Language Acquisition and Foreign Language Teaching. 2002. № 12. P. 3–33.

7. Leech G. Corpus annotation schemes // Literary and Linguistic Computing. 1993. № 8 (4). P. 275–281.

8. Моделирование в корпусной лингвистике: специализированные корпуса русского языка / В.П. Захаров, И.В. Азаров, О.А. Митрофанова [и др.]. СПб. : Изд-во СПб. ун-та, 2019. 208 с.

9. Камшилова О.Н. Анализ ошибок в корпусе ученических текстов: исследовательский и технологический аспект // Материалы Международной конференции «Корпусная лингвистика-2011», 27–29 июня 2011, Санкт-Петербург. СПб. : СПбГУ, 2011. С. 175–180.

10. Куниловская М.А. Классификация переводческих ошибок для создания разметки в учебном параллельном корпусе // Lingua Mobilis. 2013. № 1 (40). С. 141–158.

11. 赵玮. 中介语语料库词汇错误的标注问题及改进建议 // 云南师范大学学报. 2015. Vol. 13, № 2. P. 1–9.

12. Груднева Е.В., Бучилова И.А., Волкова Н.А. Корпусы ошибок: целевая аудитория, возможная архитектура корпуса // Вестник Череповецкого государственного университета. 2018. № 5 (86). С. 63–72.

doi: 10.17223/978-5-907572-04-1-2022-47

**ВЗАИМОДЕЙСТВИЕ ЭМОЦИОНАЛЬНО-ОЦЕНОЧНОЙ
ТОНАЛЬНОСТИ С ТЕМАТИКОЙ НОВОСТНЫХ ТЕКСТОВ:
ТОНАЛЬНЫЙ СЛОВАРЬ КАК СРЕДСТВО
КЛАССИФИКАЦИИ ТЕКСТОВ¹**

Петренко К.В.

Томский государственный университет, магистрант
**INTERACTION OF EMOTIONAL AND EVALUATIVE
TONALITY WITH TOPIC OF NEWS TEXTS:
TONE DICTIONARY FOR TEXT CLASSIFICATION**

Petrenko K.V.

Tomsk State University, master student

В статье охарактеризованы основные пути вычисления тональности новостного текста. Представлен пример анализа на материале

¹ Статья подготовлена в рамках проекта Министерства образования и науки Российской Федерации («Выпускник-2020»).