суу бүркүү; жумурткадан кыр чыгаруу; камырдан кыл суургандай; мурдун балта кеспеген; талпагын ташка жаюу; котур ташы койнунда; итке минген; кой оозунан чөп албаган; тилинен бал тамган; оозу менен орок оргон; бөйрөктөн шыйрак чыгарган; кулак-мурун кескендей; иттин кара капталынан; кабыргасы менен кеңешүү; текейден арзан; коендой окшош; терисин тескери сыйруу; ат терисин кургатпай; тиш каккан; баш терисин байкоо; сары изине чөп салуу и др.

Таким образом, выше речь шла в основном о частеречной разметке корпуса и принципах лематизации. Следовательно, для нас открытым остается вопрос, как учесть морфо-синтаксико-семантическую информацию в корпусе.

ГРАММАТИЧЕСКАЯ АННОТАЦИЯ ДЛЯ ГРАЛИС-КОРПУСА

Бранко Тошович
Institut für Slawistik der Karl-Franzens-Universität Graz
(Институт славистики Университета им. Карла и Франца в Граце)

1. Гралис-Корпус является параллельным корпусом для исследования славянских языков, созданным в 2007 году в Университете им. Карла и Франца в Граце [Gralis-Korpus-www]. Корпус предназначен в первую очередь для сопоставительного изучения славянских языков. По разметке Гралис-Корпус относится к аннотированным корпусам. Предусмотрены следующие типы разметок: метаязыковая, морфологическая, синтаксическая, семантическая и стилистическая, две из которых (метаязыковая и морфологическая) уже существуют. Морфологическая аннотация проведена для сербского, хорватского и бошняцкого языков при помощи аналитическо-синтетической системы «MorphoGenerator».

2. Работа над морфологической аннотацией началась весной 2008 года в рамках проекта «Различия между боснийским/бошняцким, сербским и хорватским языками» (2006–2010) австрийского Фонда для поддержки научных исследований и продолжилась в рамках проекта «Lexikarium», который финансировало Правительство Штирии (2008–2013). Систему аннотации и процесс порождения форм и парадигм разработал Бранко Тошович, а программную оболочку Ольга Ленер.

3. Так как Гралис-Корпус охватывает ряд языков, необходимо было выбрать ту грамматическую разметку, которая бы больше всего годилась для их сопоставления и изучения. Такой нам показалась Multext-East кодировка (Multilingual Texts and Corpora for Eastern and Central European Languages – multilingual dataset for language engineering research and development: MultiText East-www), разработанная в 2004 году группой авторов во главе с Томажем Ерявцем, так она являлась унифицированной для большинства славянских языков. Для поиска в

Гралис-Корпусе на основе грамматической разметки выбран CQP-синтаксис, представляющий очень широкие и разнообразные возможности.

4. Грамматическая кодировка для Гралис-Корпуса состоит из следующих позиций: 1 – часть речи, 2 – подтип части речи, 3 – наклонение, 4 – время, 5 – лицо, 6 – число, 7 – род, 8 – залог, 9 – отрицание, 10 – определенность, 11 – возвратность, 12 – падеж, 13 – одушевленность, 14 – клитика, 15 – вид, 16 – вежливость, 17 – переходность, 18 – ударение, 19 – способ глагольного действия, 20 – лексико-семантическая группа, 21 – сочетание (управление, согласование), 22 – экспрессия, 23 – функциональный стиль, 24 – аналитическая форма, 25 – деструкция, 26 – словообразование, 27 – номер порождения, 28 – грамматический тип порождения.

На основе этой общей модели создана конкретная схема с 20-ю позициями, охватывающая все части речи и их категории. Для обозначения таксонов в каждой позиции используются малые латинские буквы (лишь для указания на лицо и тип парадигмы даются цифры). Так как позиций больше, чем букв, некоторые из графем повторяются, но путаница исключается, так как первую позицию всегда занимает название части речи, что предотвращает любую двусмысленность:

1. **Часть речи**: **n –** существительные, **v –** глаголы, **a –** прилагательные, **p –** местоимения, **r** – наречия, **s –** предлоги, **c –** союзы, **m –** числительные, **i –** междометия, **q –** частицы, **y –** аббревиатуры

2. **Подтип части речи:** Существительные: **c** – нариц., **p** – собств., **m** – веществ., **l** – собират. Глаголы: **m** – полнозначн., **a** – вспомогат., **o** – модальн., **c** – связочн., **b** – базовый. Прилагательные: **f** – качеств., **r** – относит., **m** – веществ., **s** – притяж. , **o** – порядк., **m** – количеств. Местоимения: **p** – личн., **d** – указат., **i** – неопред., **s** – притяж., **q** – вопрос., **r** – относит., **x** – возвр., **z** – отриц., **g** – общ., **y** – вопрос.- относ., **j** – определ., **t** – указат.-относит. Наречия: **g** – общ., **z** – отриц., **a** – адъективн., **v** – глагольн., **q** – вопросит. Предлоги: **p** – препозитивн., **t** – постпозитивн. Союзы: **c** – сочин., **s** – подчинит. Числительные: **c** – количеств., **o** – порядк., **m** – итеративн., **l** – видов., **s** – специальн. Частицы: **z** – отриц., **q** – вопрос., **o** – модальн., **r** – положит. Аббревиатуры: **n** – именн., r – наречн.

3. **Тип формы:** Глаголы: **i** – изъявит., накл., **m** – повелит. накл., **c** – сосл. накл. 1, h – сосл. накл. 2, **n** – инфинитив, **p** – причастие, **g** – деепричастие 1 (несов. в. / нас. вр.), **w** – деепричастие 2 (сов. в. / прош. вр.), **u** – супин, **t** – переходн., **q** – циритов., **s** – гипотетич. Прилагательные: степени сравнения – **p**: положит. ст., **c** – сравнит. ст., **s** – превосх. ст. Наречия: степени сравнения – **p**: положит. ст., **c** – сравнит. ст., **s** – превосх. ст., **e** – элятив.

4. **Время**: **p** – презенс, **i** – имперфект, **f** – будущ. 1 Ср (серб.), **w** – будущ. 1 Хр (хорв.) , **z** – будущ. 1 Ср/Хр (серб./хорв.), **q** – будущ. 2, **s** – перфект, **l** – плюсквамперфект 1, **t** – плюсквамперфект 2, **a** – аорист

5. **Лицо**: **1** – первое л., **2** – второе л., 3 – третье л.

6. **Число**: **s** – ед. ч., **p** – мн. ч., **d** – двойств. ч., **l** – собират. ч.

7. **Род**: **m** – м. р., **f** – ж. р., **n** – ср. р., **l** – общий р.

8. **Залог**: **a** – действ. залог, **p** – страд. залог

9. **Отрицание**: **n** – да, **y –** нет

10. **Определенность**: **n** – да, **y –** нет

11. **Возвратность: n** – да, **y –** нет

12. **Падеж**: **n** – им. п., **g** – род. п., **d** – дат. п., **a** – вин. п. **v** – зват. п., **l** – предл. п., **i** – творит. п.

13. **Одушевленность**: **n** – да, **y –** нет

14. **Клитика**: **n** – да, **y –** нет

15. **Вид**: **p** – несов. в., **e** – сов. в., **b** – двойн. в.

16. **Вежливость**: **n** – да, **y –** нет

17. **Переходность**: **n** – да, **y –** нет

18. **Деструкция**: **n** – да, **y –** нет

19. **Словообразование**: **s** – непроизв., **c** – сложн.

20. **Тип**: 01, 02, 03…

Незаполненными позициями пока являются: 11. возвратность, 16. вежливость, 17. переходность, 18. деструкция, 19. словообразование. Из 28 запланированных позиций отсутствуют следующие: 18. ударение (которое дается в «Akzentarium-е»1), 19. способ глагольного действия, 20. лексико-семантическая группа, 21. сочетание (управление, согласование), 22. экспрессия, 23. функциональный стиль, 24. аналитическая форма.

5. При помощи «MorphoGenerator-а» можно получать парадигму любого слова, что особенно важно в процессе обучения. Межъязыковые различия в частоте использования словоформ между языками не определяются произвольно, в общих формулировках, а объективно при помощи корпусных данных. «Gralis-MorphoGenerator» предоставляет статистическую информацию о каждой части речи в Гралис-Корпусе.

6. Морфологическая разметка для сербского, хорватского, бошняцкого и черногорского языков проведена в настоящее время для 100 461 слова. Их парадигмы выявлены на основе 822 правил[2]. Процедура разметки выглядит следующим образом. **а)** Создается список всех слов в рамках одного литературного языка (в случае Гралис-разметки их три: сербский, хорватский и бошняцкий). б) Слова разделяются по частям речи. **в)** В рамках каждой части речи слова объединяются в отдельный список по грамматическому признаку, важному для порождения парадигмы (для существительных это категория рода, для прилагательных наличие/отсутствие степеней сравнения, для глаголов вид). **г)** Полученный спи-

---

[1] Существует прямая связь между «MorphoGenerator-ом» и «Akzentarium-ом» – базой данных с ударениями (система ударений сербского, хорватского и бошняцкого языков является очень сложной: она представляет собой сочетание двух просодических характеристик – долготы, изменения тона, подъема и снижения, на основе чего выделяются четыре ударения: долгое восходящее, краткое восходящее, долгое нисходящее, краткое нисходящее, а также заударная долгота).

[2] Под правилом подразумевается матрица соотношений между кодовыми позициями и парадигматическими заполнениями, точнее, процедура раскрытия системы изменения слов, их варьирования и представления в целях порождения всех форм и целостной парадигмы.

сок расчленяется на подсписки, учитывая дополнительный грамматический признак (для существительных это одушевленность, для глаголов переходность). Этим заканчивается подготовка словарного материала для его грамматической обработки **д).** Проводится анализ списка под г и выявляются парадигматические маркеры – падежи, лица, являющиеся релевантными для порождения целостной парадигмы (напр., для существительных это им., род., зв. и тв. п. ед. ч. и им. и род. п. мн. ч.). **е)** Создается список слов с общими грамматическими маркерами (скажем, для существительных мужского рода на согласный, имеющих в им. п. ед. ч. беглую гласную **а**. **ж)** Выбирается типичное слово в этой группе и для него создается таблица со всеми 20-ю позициями и всеми словоформами. По этом слову и номеру типа называется каждое правило (напр. 174}$^{\text{mudrac}}$). Этим заканчивается лингвистическая работа и начинается программная. **з)** Программист берет материал, полученный в процедурах **а** – **ж,** и превращает его в реляционную базу данных MySQL, распределяя слова не по окончаниям, а по конечным буквам. **и)** Создаются две маски – одна для поиска информации по грамматической разметке и вторая для отображения результатов поиска. Обе они объединены в MorphoGenerator. **к)** Он соотносится с Гралис-Корпусом и создаются информационные каналы в двух направлениях – от MorphoGenerator-а к Корпусу и от Корпуса к MorphoGenerator-у, для того, чтобы любую форму, приведенную в MorphoGenerator-е, можно было найти в Корпусе и получить информацию о ее частоте, и чтобы для любого словоупотребления в Корпусе можно было получить парадигму в MorphoGenerator-е.

7. Проведенная морфологическая разметка указала на следующие закономерности. Во-первых, существуют два противоположных правила порождения форм и парадигм: а) правила, при помощи которых можно получить парадигму только для одного слова, б) правила для порождения парадигм для сотни и тысячи слов. Во-вторых, выделяются общие и частичные правила. Первые касаются типа изменения (склонения, спряжения, сравнения), вторые относятся к некоторым грамматическим категориям. В-третьих, различаются структурные, категориальные и интеркатегориальные правила. Структурные правила указывают на то, какие формальные средства используются для порождения парадигм и какие чередования происходят на стыке окончания и предшествующей морфемы. Категориальные правила покрывают морфологические своеобразия в рамках определенной грамматической категории (рода, числа, вида, залога и т. п.). В-четвертых, на появление большого числа правил не влияет большое число окончаний, о чем свидетельствует и тот факт, что в порождении форм существительных участвует всего 11 окончаний **(-а**, **-о**, **-е**, **-i**, **-u**, **-ø**, **-oj**, **-om**, **-em**, **-ama**, **-ima**), которые повторяются в виде синкретизма. В-пятых, число и структура правил зависят от каждой части речи. Больше всего правил требуют глаголы (378), затем существительные (311), намного меньше прилагательные (71) и местоимения (50), а на последнем месте находятся числительные (12). Однако, если число правил соотнести с числом слов, то получается совсем иная картина: 135 местоимений нуждаются в 50 правилах, 198 числительных в 12, в то время как для 37 606 существительных

надо 311, для 30 030 глаголов 378, для 32 492 прилагательных 71. По отношению числа слов и числа правил самый низкий параметр наблюдается у прилагательных (0,0022) и существительных (0,0083), более высокий у местоимений (0,37), а самый высокий у числительных (0,6%). Для аннотации 37 606 существительных надо иметь 311 правил (почти нет существительных без склонения), которые являются сложными из-за разнообразия грамматических категорий (семь падежей, два числа, три рода, категории одушевленности), разнообразных чередований и наличия вариантных форм. Парадигмы прилагательных (32 492) порождаются при помощи 71 правила и охватывают категории, характерные для существительных (род, число, падеж), а также категорию сравнения и определенности. Местоимения составляют закрытое множество, состоящее из 135 единиц. Для порождения их парадигм необходимо 50 правил. Для еще одной части речи, образующей закрытое множество, – числительных (198), нужно иметь 12 правил. Что касается глаголов, их правила являются самыми сложными, так как охватывают больше категорий, чем другие части речи (наклонение, вид, лицо, залог, род, спряжение, склонение, переходность, возвратность, /не/одушевленность и др.), обладают широкой системой чередований и вариативностью. Поэтому 30 030 глаголов покрывает 378 правил. В-шестых, создание и существование трех отдельных норм для трех очень близких языков – сербского, хорватского и бошняцкого (порог понимания между их носителями является в повседневной коммуникации почти стопроцентным) на одной и той же диалектной базе – штокавской требуют учитывать все нормативные различия, которых не так много, но которые все-таки усложняют морфологическую разметку. Из-за этого своеобразия грамматическая аннотация для Гралис-Корпуса является уникальной (нам неизвестны случаи, когда разметка охватывает одновременно несколько языков).

8. Что касается дальнейшей работы, она будет продолжена в трех направлениях. **(1)** Так как в период с 2008 по 2013 г. была проведена морфологическая разметка 100 461 изменяемого слова и около 11 000 неизменяемых, что составляет приблизительно 112 000 слов, в 2014 году будет проверена созданная система, особенно правила и результаты их порождения, проведена корректура всего того, что оказалось неточным, слабым, неясным, после чего будет открыт Gralis-MorphoGenerator для всех желающих. В 2015 году основное внимание будет уделено снятию грамматической омонимии. **(2)** В 2016 году в центре внимания окажется морфологическое аннотирование других славянских языков (в этих целях предусмотрена широкая консультация со специалистами из других стран, чтобы существующие системы для разметки отдельных языков модифицировать для Гралис-Корпуса или, в случае совместимости, полностью принять). **(3)** В 2017 году начнется работа над моделью синтаксической, семантической и стилистической разметки.

Литература

1. Gralis-Korpus // http://www-gewi.uni-graz.at/gralis/korpusarium/gralis_korpus.html. Состояние: 10. 10. 2013.

2. *Tošović, Branko (Hg)*. Das Gralis-Korpus // Branko Tošović (Hg.). *Die Unterschiede zwischen dem Bosnischen/Bosniakischen, Kroatischen und Serbischen.* 2008. Graz. 724–749.

3. *Tošović, Branko.* Гралис-Корпус // *Wiener slawischer Almanach.* 2013. München, 83. 89–111.

4. Т*ошович, Бранко.* Сопоставительное изучение славянских языков при помощи многоязычного «Гралис-Корпуса» // *Izučavanje slovenskih jezika, književnosti i kultura kao inoslovenskih i stranih.* – Beograd: Slavističko društvo Srbije. 2008. 336–340.

К ПРОБЛЕМЕ УНИФИКАЦИИ СИСТЕМЫ ОБОЗНАЧЕНИЯ ГРАММАТИЧЕСКИХ КАТЕГОРИЙ В КОРПУСАХ ТЮРКСКИХ ЯЗЫКОВ

Б.Э. Хакимов, А.М. Галиева, А.Р. Гатиатуллин
Казанский федеральный университет, НИИ «Прикладная семиотика» АН РТ

В статье представлены первые итоги сравнительного анализа систем грамматической разметки в различных корпусах тюркских языков. Анализируются именное и глагольное словоизменение, а также метаязык формального описания грамматических категорий.

1. Введение

Развитие тюркологии последних лет отмечено углублением теоретической базы лингвистических исследований, повышением внимания к новым направлениям и проблемам современного языкознания, в том числе и прикладного. Критерии и принципы содержательного анализа словоизменительных категорий остаются одним из приоритетных направлений исследований по тюркским языкам, при этом интерес к этой традиционной тематике поддерживается постановкой новых задач, обусловленных развитием информационных технологий.

Представление в корпусной аннотации информации о грамматических категориях тюркских языков, как показывает опыт разработчиков, является самостоятельной научной проблемой, пересекающейся с разными, порой противоположными подходами к описанию грамматических явлений.

Одна из важнейших задач разработки грамматической аннотации в корпусах тюркских языков – выявить фондовый (инвентарный) уровень словоизменительных категорий и создать оптимальный метаязык описания этих грамматических категорий.

2. Общие принципы

Как известно, тюркское языкознание в течение продолжительного времени развивалось под сильным и непосредственным влиянием индоевропейского и русского языкознания, когда отдельные факты и даже грамматические категории