НЕКОТОРЫЕ ОСОБЕННОСТИ АДЫГЕЙСКОГО ЭЛЕКТРОННОГО КОРПУСА

Юрий Александрович Ландер, к.ф.н.
НИУ «Высшая школа экономики», Институт востоковедения РАН
E-mail: yulander@yandex.ru

Тимофей Александрович Архангельский, к.ф.н.
Universität Hamburg
E-mail: timarkh@gmail.com

Ирина Гаруновна Багирокова
Институт языкознания РАН, НИУ «Высшая школа экономики»
E-mail: ibagirokova@yandex.ru

Аннотация: В работе описываются формальные и концептуальные решения, принятые в разрабатываемом авторами электронном корпусе адыгейских текстов. Рассматриваемые решения связаны с тем, как производится автоматический морфологический разбор адыгейских текстов, требуемый для грамматической аннотации морфем и словоформ.

Ключевые слова: Адыгейский корпус, адыгейский язык, корпусная лингвистика

SOME SPECIFIC FEATURES OF THE WEST CIRCASSIAN ELECTRONIC CORPUS

Yury Lander
National Research University Higher School of Economics, Institute of Oriental Studies RAS
E-mail: yulander@yandex.ru

Timofey Arkhangelskiy
Universität Hamburg
E-mail: timofey.arkhangelskiy@uni-hamburg.de

Irina Bagirokova
Institute of Linguistics RAS, National Research University Higher School of Economics
E-mail: ibagirokova@yandex.ru

Abstract: The paper deals with concepts and formal conventions which are accepted for the West Circassian electronic corpus by its authors. The ideas discussed here are induced by the specifics of automatic morphological parsing of West Circassian texts required for grammatical annotation of morphemes and word forms.

Keywords: West Circassian corpus, West Circassian, corpus linguistics

1. Введение

Языковые корпуса – электронные коллекции текстов, сопровождаемых разметкой, которая позволяет проводить автоматический поиск по имеющемуся материалу, – за короткое время стали неотъемлемой частью лингвистики. Популярностью корпуса обязаны тому, что благодаря им можно проследить, как язык используется в реальности. Это противопоставляет корпусные данные данным, полученным в результате интроспеции или экспериментов, и делает их полезными не только для лингвистов, но и для педагогов, литературоведов и фольклористов, историков: при условии, что в корпусе представлен достаточный материал, с помощью него можно наблюдать вариативность в языке в зависимости от времени, особенности стиля авторов и текстов, определять более и менее употребимые выражения.

Особенно важными для решения этих задач оказываются, во-первых, объем корпуса, а во-вторых, представленность в нем разнообразных текстов. Не удивительно, что в первую очередь корпуса создаются для наиболее крупных языков – как из-за популярности их исследования, так и благодаря тому, что для них имеется больше доступного материала. Естественно, и архитектура корпусов фактически подстраивалась именно под наиболее крупные языки – прежде всего, под европейские.

Абхазо-адыгские языки находятся в этом отношении в двояком положении. С одной стороны, живые абхазо-адыгские языки обладают некоторой литературной традицией, что уже позволяет говорить о задокументированности письменных языков и делает возможным создание соответствующих корпусов. С другой стороны, типологически эти языки значительно отличаются от языков, которые оказали влияние на обычные представления о том,

как должен выглядеть корпус. Несмотря на это, а отчасти и благодаря этому (ибо создание корпуса такого языка фактически стало вызовом для мировой корпусной лингвистики), в 2014 году началась работа над адыгейским корпусом, снабженным грамматической разметкой, которая является результатом автоматического (!) анализа словоформ – и тем самым может быть получена для большого количества текстов.

Основой анализа адыгейских текстов стал подход к адыгейскому языку, разрабатываемый в рамках проекта исследования абхазо-адыгских языков в рамках экспедиций, которые организовывались и организуются Российским государственным гуманитарным университетом и НИУ «Высшая школа экономики». Этот подход, отчасти, но не в окончательном виде отраженный в сборнике [3], во многом основывался на уже имеющихся детальных описаниях адыгейского языка (прежде всего, [11; 20]), и на работах по внутригрупповому сравнению вроде [4; 6; 7; 8; 9; 12; 16]. Одновременно участники этого проекта пытались учитывать и типологические запросы к материалу, что обусловливало его специфику. Ниже мы опишем некоторые особенности Адыгейского корпуса, которые необходимо учитывать при его использовании.

2. Принципы представления морфологической информации

Как и в других полисинтетических языках, морфология в адыгейском языке играет намного более важную роль в построении высказывания, чем в неполисинтетических европейских языках. Не случайно, в Адыгейском корпусе информация о морфологической структуре (например, наличие и расположение тех или иных морфем и их комбинаций) оказывается крайне существенной для учета и поиска словоформ (см. подробнее [13]). Для отображения этой информации в адыгейском языке, в котором за редкими исключениями слова четко делятся на морфемы, в Адыгейском корпусе используется система морфологического глоссирования, которая следует основным принципам морфологической нотации, принятым в типологии (см., например, [18; 19]). Вкратце эта система может быть представлена следующим образом. Слова делятся на морфемы, разделяемые дефисом, и такой последовательности морфем приводится соответствие последовательности семантически/функционально мотивированных условных обозначений, данных этим морфемам, как в следующем примере:

къы-т-фы-тыр-а-р-и-гъэ-лъ-хьа-гъ

DIR-1PL.IO-BEN-LOC-3PL.IO-DAT-3SG.ERG-CAUS-лежать-уносить-PST

Он заставил их положить это туда для нас

Условное представление в виде глоссирования позволяет, в частности, искать морфемы и их комбинации, отвлекаясь от проблемы существования алломорфов (в адыгейском языке, как правило, обусловленных морфонологически), а также объединять морфемы в классы (например, любые личные префиксы непрямого объекта могут быть найдены без уточнения лично-числовых характеристик как PERS.IO). Кроме того, такое представление делает корпус более доступным для типологов, знакомых с общем строем адыгейского языка, но не с конкретными его морфемами.

При этом, однако, сохраняется проблема выделения морфем и их значений: в разных работах могут быть приняты разные решения по этому поводу. Мы руководствуемся в первую очередь **принципом формального обоснования**: морфемы выделяются на основе их формальных свойств. Соответственно, мы почти (!) не учитываем при выделении и противопоставлении морфем их семантику. В то же время нами учитываются морфонологические свойства и сочетаемость морфем, что, например, позволяет противопоставить суффикс номинализации образа действия творительному падежу (ср. *я-гупшыса-кlэ* 3PL.P+POSS-думать-NMZ.MNR ‘образ мыслей’ vs *я-гупшысэ-кlэ* 3PL.P+POSS-думать-INS ‘их мыслями’; для последней формы, впрочем, разбор с номинализацией тоже допустим в определенных контекстах). Основная мотивация следования принципу формального обоснования – это то, что поморфемный разбор словоформ производится автоматически, а формальную сочетаемость морфем задать правилами проще, чем семантическую. Это приобретает особое значение для языков типа адыгейского, где многие морфемы обладают исключительно широкой сочетаемостью. Многочисленные следствия принципа формального обоснования мы увидим ниже.

3. Особенности терминологии

Принятая в корпусе терминология, в том числе терминология, которая лежит в основе используемых нами глосс и используется при запросах, может быть не всегда привычна отечественному читателю.

Типичный пример – названия падежей. Для показателей двух ядерных адыгейских падежей используются сокращения ABS (абсолютивный падеж – соответствует

именительному падежу в традиционных описаниях) и OBL (косвенный падеж – соответствует эргативному падежу в традиционных описаниях). И в том и в другом случае выбор оправдан типологически. Термин «именительный падеж» плохо применим к показателю *-р* уже потому, что этот маркер не используется при обозначении наименования; термин «абсолютив» является стандартным при описании эргативных языков (см., например, [14]). Термин «эргативный падеж» неудобен, так как он ориентируется только на одну функцию соответствующего падежа – маркирование агенса при переходном глаголе, в то время как этот падеж используется вообще при наличии соответствующих лично-числовых префиксов где-либо – в том числе при маркировании непрямого объекта, посессора, а иногда и абсолютивного актанта [2, с. 80–83]. Вместо термина «эргативный падеж» используется понятие косвенного падежа – как второго ядерного падежа, особенно в двухпадежных системах [1], к которым близка адыгейская. Для глоссирования показателей периферийных падежей принимаются обозначения INS («инструменталис» - для творительного падежа) и ADV («адвербиалис» - для так называемого превратительного падежа). Если первая глосса навряд ли должна вызывать сложности, вторая оправдана скорее тем, что автоматическое глоссирование рассматривает морфему *-эу* как единую независимо от того, маркирует ли она именные группы в функции «вторичной предикации» [21], образует ли наречные или деепричастные выражения, и нам неизвестны формальные морфологические основания для противопоставления этих функций.

Проблема терминологических различий, естественно, не является принципиальной, если она сводится к простому переводу из одной системы терминов в другую. Ниже, однако, мы коснемся нескольких более сложных вопросов, для которых приходилось принимать концептуальные решения.

4. Некоторые концептуальные особенности

Личные префиксы. В корпусе противопоставляется абсолютивная серия личных префиксов (PERS.ABS), серия личных префиксов непрямого объекта (PERS.IO) и серия эргативных личных префиксов (PERS.ERG). Важно, что в отличие от многих описаний, мы избегаем использование терминов вроде «субъект»/«подлежащее», «прямое дополнение», поскольку считаем их в большей степени ориентированным не на форму, а на функцию. Абсолютивная серия в литературном адыгейском обычно отличается огласовкой *-ы*, располагается в начале словоформы и может представлять как то, что переводится на русский язык подлежащим (прежде всего, в отсутствие эргативного личного префикса), как и то, что переводится на русский язык прямым дополнением (обычно при наличии эргативного

префикса), хотя собственно перевод, разумеется, не должен считаться мерилом грамматической интерпретации. Специально отметим, что префикс *мэ-* в корпусе трактуется не как личный префикс, а как показатель динамичности (ср. [11]), хотя в литературе встречаются и другие трактовки [9; 16, с. 40].

Аппликативы. Конструкции с непрямыми объектами, как правило, включают помимо личных префиксов соответствующей серии (формально отсутствующих при 3 лице единственного числа и не отражаемых при глоссировании), показатели, до некоторой степени специфицирующие семантическую роль непрямого объекта, - аппликативы (показатели, вводящие дополнительного участника в качестве актанта; см. обсуждение в [2; 10; 17]). В традиционном адыговедении (см., например, [11]) подобные показатели фактически описываются в нескольких разделах: категория объектной версии (*с-а-ф-и-гъэ-хьазыры-гъ* 1SG.ABS-3PL.IO-BEN-3SG.ERG-CAUS-готовый-PST ‘он меня к ним подготовил’), категория союзности (*а-дэ-з-гощы-щт* 3PL.IO-COM-3SG.ERG-делить-FUT ‘я это с ними разделю), локативные превербы (*а-тыр-и-гъэ-ты-щты-гъ* 3PL.IO-LOC-LOC-CAUS-стоять-AUX-PST ‘он заставлял его стоять на них’), категории префиксального потенциалиса (*къы-с-фэ-шlы-щт-эп* DIR-1SG.IO-BEN-делать-FUT-NEG ‘я не смогу это сделать’) и непроизвольности (*lэклэ-шlы-клы-гъ* INADV-делать-EL-PST ‘он случайно сделал это’), категория отторжимой принадлежности (см. ниже). Традиционная категория версии в корпусе представляется как появление бенефактива (*къ-а-ф-и-гъэ-хьы-гъ* DIR-3PL.IO-BEN-3SG.ERG-CAUS-нести-PST ‘он им это прислал’), малефактива (*къы-з-шlо-з-гъэ-шlы-гъ* DIR-REL.IO/RFL.IO-MAL-1SG.ERG-CAUS-делать-PST ‘я представил себе это’) или общелокативного преверба (*сы-п-щы-щына-гъ* 1SG.ABS-2SG.IO-GENLOC-бояться-PST ‘я тебя испугался’). Префиксальный потенциалис трактуется в корпусе по формальным признакам как разновидность бенефактивной конструкции, обычно описываемой как категория версии. Прочие категории противопоставляются, хотя локативные превербы не специфицируются по значению и всегда глоссируются как LOC. Существенно, однако, что есть возможность поиска аппликативов без спецификации их функции (APP). Например, запрос APP-PERS.ERG-*-PST# (последовательность аппликатива, эргативного личного префикса, любых морфем и конечного показателя прошедшего времени) дает результаты с бенефактивным префиксом (*фэ-с-lотэ-жьы-гъа-гъ* BEN-1SG.ERG-рассказать-RE-PST-PST ‘я ему это пересказывал’), с локативными превербами (*щы-лъ-и-гъэ-кlота-гъ* GENLOC-LOC-3SG.ERG-CAUS-переместиться-PST ‘он передвинул это там’) и т.д.

Дативная конструкция. В трактовке корпуса, следующей, например, [11; 17] в ряде адыгейских глаголов возникает особый префикс (*й*)э-, глоссируемый как DAT («дативный»). Этот показатель является аппликативом, хотя и не специфицирует роль непрямого объекта, выводя ее из семантики основы (*къы-с-э-упчlы-гъ* DIR-1SG.IO-DAT-спросить-PST ‘он спросил меня’). В 3 лице единственного числа личный префикс непрямого объекта отсутствует, так что при основе появляется исключительно дативный преверб. Таким образом, в корпусе не постулируется особое выражение непрямого объекта 3 лица единственного числа *е-*, как это иногда делается (но см. [11], где *е-*, как и у нас, трактуется как префикс косвенного отношения). Заметим, однако, что дативная конструкция – единственная, в которой аппликатив (дативный префикс) может исчезать по морфонологическим причинам (*зы-с-а-гъэ-фэпа-гъ* RFL.ABS-1SG.IO-3PL.ERG-CAUS-одеть-PST ‘они меня заставили одеться’).

Посессивные конструкции. В притяжательных (посессивных) формах имен морфема *и-* трактуется единообразно как посессивный аппликатив, вводящий личный префикс посессора (*у-и-гухэлъы-шlу-хэ-р* 2SG.P-POSS-намерение-добрый-PL-ABS ‘твои добрые намерения’). Хотя такие личные префиксы, как и другие личные префиксы, вводимые аппликативами, принадлежат серии непрямого объекта, для удобства пользователя в корпусе они на данный момент трактуются как представители серии посессора (PERS.P). В результате они не вполне правомерно объединяются с префиксами неотторжимого посессора, появляющимися без префикса *и-*. Тот же посессивный *и-* постулируется в глаголе ‘иметь / быть у кого-то’ *и-lэ-* с корнем ‘быть’ и в образованиях с корнем *е-* ‘иметься’. Соответственно, он противопоставляется омонимичному локативному превербу *и-*; формальным основанием для этого является, в частности, их разное поведение при личном префиксе 3 лица множественного числа; ср. *я-lэ-х* 3PL.P+POSS-быть-PL ‘они у них есть’ vs *а-ры-сы-х* (< *а-и-сы-х*) 3PL.IO-LOC-сидеть-PL ‘они в них сидят’.

Сложные показатели. В силу принципа формальной обоснованности некоторые традиционно выделяемые показатели в корпусе представляются как сочетания морфем – независимо от того, можно ли приписать им единое значение. Таковы, например, показатели времени вроде давнопрошедшего *-гъагъэ*, который трактуется в корпусе как сочетание двух показателей прошедшего времени *-гъэ* (тем более, что они иногда могут разрываться, как в *къы-зэ-те-уцо-гъэ-па-гъ* DIR-REC.IO-LOC-встать-PST-ASRT-PST ‘он окончательно встал на ноги’), и сложные превербы (*клэ-лъы-плъэ-зэ* LOC-LOC-смотреть-SIM ‘наблюдая за ним’). Исключением является так называемый показатель причастий образа действия и фактивных причастий *зэрэ-* (условно глоссируемый REL.SUB), который в силу устоявшейся традиции

трактуется как единая морфема, хотя, в принципе, является членимым – синхронно или диахронически [4; 5].

***Омонимия показателей* -щт**. Многие временные показатели содержат в качестве одной из частей элемент *-щт-*. Мы в этой связи противопоставляем показатель будущего времени FUT (сочетающийся, однако, с другими временными показателями – например, в формах вроде *я-лъыты-гъэ-щт* 3PL.ERG-считать-PST-FUT ‘это будет зависеть от них’) и вспомогательный морфологический элемент – очевидно, превратившийся в аффикс вспомогательный глагол AUX, к которому далее присоединяются показатели времени, как в формах вроде *къ-ы-уаты-щты-гъа-гъ-эп* ‘они бы тебе это не дали’. Это противопоставление имеет морфонологические основания [15], то есть тоже обосновывается формально.

***Показатель* -н**. Суффикс *-н* выступает в качестве показателя масдара (*къэ-гъэ-лъэгъо-ны-р* DIR-CAUS-быть.видимым-MOD-ABS ‘показ, демонстрация’), а также одного из будущих времен (*п-щэ-н-х-а* 2SG.ERG-вести-MOD-PL-Q ‘ты отведешь их?’) и при выражении эпистемической модальности (*макӀ-эпы-н* (*фае*) мало-NEG-MOD ‘должно быть, не мало’). Но хотя эти функции, вероятно, полезно разграничивать – в том числе на основе сочетаемости образуемых форм с падежами, позиции суффикса в словоформе и т.д., во множестве форм принять решение об их разграничении на основе формальной информации невозможно (ср. формы типа *кӀо-н* идти-MOD ‘хождение; он пойдет’). Поскольку значения *-н*, по-видимому, являются связанными, а постулирование нескольких суффиксов в данном случае приводит к огромному количеству омонимии при автоматическом разборе, в корпусе этот суффикс всегда глоссируется единообразно как модальный показатель MOD.

5. Заключение

В этой работе мы рассмотрели формальные и концептуальные особенности разрабатываемого нами корпуса адыгейских текстов. Важнейшей характеристикой этого корпуса является морфологическая разметка, связанная с делением словоформ на морфемы и приписыванием морфемам условных ярлыков (глосс). Выше мы эксплицировали некоторые решения, принятые для Адыгейского корпуса, которые пользователь должен учитывать при работе с корпусом.

Как показано выше, при выделении и отождествлении или дифференциации разных морфем мы руководствовались в первую очередь не функциональными и семантическими характеристиками, а формальными свойствами наблюдаемых единиц. Это обусловлено тем, что разметка текстов, включаемых в корпус, проводится автоматически, а кроме того,

использование корпуса, как кажется, должно подразумевать такую его архитектуру, которая наименее зависима от конкретных теорий (хотя, как мы видели, целиком избежать подобной зависимости, вероятно, невозможно).

Литература

1. Аркадьев П.М. Функционально-семантическая типология двухпадежных систем // Вопросы языкознания. 2005. № 4. С. 101–120.

2. Аркадьев П.М., Аркадьев П.М., Ландер Ю.А., Летучий А.Б., Сумбатова Н.Р., Тестелец Я.Г. 2009. Введение. Основные сведения об адыгейском языке // Аспекты полисинтетизма: очерки по грамматике адыгейского языка – М.: РГГУ, 2009. – С. 17–120.

3. Аспекты полисинтетизма: очерки по грамматике адыгейского языка / Под ред. Я.Г. Тестельца и др. М.: РГГУ, 2009.

4. Бижоев Б.Ч. Причастие в адыгских языках в сравнительном освещении. Нальчик: Нарт, 1991.

5. Герасимов Д.В., Ландер Ю.А. 2008. Релятивизация под маской номинализации и фактивный аргумент в адыгейском языке // Исследования по отглагольной деривации. – М.: Языки славянских культур, 2008. – С. 290–313.

6. Гишев Н.Т. Сравнительный анализ адыгских языков. Майкоп: Качество, 2003.

7. Кумахов М.А. Морфология адыгских языков. Синхронно-диахронная характеристика. I. Введение, структура слова, словообразование частей речи. Нальчик: Кабардино-Балкарское книжное изд., 1964.

8. Кумахов М.А. Словоизменение адыгских языков. М.: Наука, 1971.

9. Кумахов М.А. Сравнительно-историческая грамматика адыгских (черкесских) языков. М.: Наука, 1989.

10. Ландер Ю.А. Актанты и сирконстанты в морфологии и в синтаксисе адыгейского языка // Вестник Российского государственного гуманитарного университета. Серия: История. Филология. Культурология. Востоковедение. 2015. № 1. С. 7–31.

11. Рогава Г.В., Керашева З.И. Грамматика адыгейского языка. Краснодар, Майкоп: Краснодарское книжное изд., 1966.

12. Урусов Х.Ш. Морфемика адыгских языков. Нальчик: Эльбрус, 1980ю

13. Arkhangelskiy T., Lander Yu. Developing a polysynthetic language corpus: problems and solutions // Диалог 2016 / Компьютерная лингвистика и интеллектуальные технологии. 2016. No. 15 (22). С. 38–47.

14. Dixon R.M.W. Ergativity. Cambridge: Cambridge University Press, 1994.

15. Korotkova N., Lander Yu. Deriving affix order in polysynthesis: evidence from Adyghe // Morphology. 2010. Vol. 20, No. 2. P. 299–319

16. Kumakhov M., Vamling K. Circassian Clause Structure. Malmö: Malmö University, 2009.

17. Lander Yu., Letuchiy A. Valency-decreasing operations in a valency-increasing language? // Verb Valency Change: Theoretical and Typological Perspectives / Ed. by A. Álvarez González, Ì. Navarro. – Amsterdam: John Benjamins. – P. 286–304.

18. Lehmann Chr. Directions for interlinear morphemic translations // *Folia Linguistica.* 1982. Vol. 16. P. 199–224.

19. Leipzig Glossing Rules. 2008. URL: https://www.eva.mpg.de/lingua/resources/glossing-rules.php (дата обращения 20.09.2018).

20. Smeets R. Studies in West Circassian Phonology and Morphology. Leiden: The Hakuchi Press, 1984.

21. Vydrin A. Are there depictives in Adyghe? // Secondary Predicates in Eastern European Languages and Beyond / Ed. by Chr. Schroeder, G. Hentschel, W. Boeder. – Oldenburg: BIS, 2008. – P. 424–445.