АДЫГЕЙСКИЙ КОРПУС И ОРФОГРАФИЧЕСКОЕ СЛОВО

Ю.А. Ландер (НИУ «Высшая школа экономики» / Институт востоковедения РАН)
Т.А. Архангельский (НИУ «Высшая школа экономики»)

1. Известно, что в полисинтетических языках морфология берет на себя ряд функций, которые в языках неполисинтетических обычно приписываются синтаксису. Этот факт настолько важен для этих языков, что некоторыми исследователями он даже берется в качестве основного, мотивирующего появление особого полисинтетического типа; см. обсуждение подходов к полисинтетизму в [Ландер 2011].

Естественным результатом «передачи» некоторых синтаксических функций морфологии оказывается некоторая синтаксичность последней [de Reuse 2009] и, как следствие, определенная размытость границ между морфологией и синтаксисом. (Разумеется, наша постановка вопроса основывается на заведомо сомнительной точке зрения, согласно которой языки среднеевропейского стандарта отражают некое каноническое противопоставление морфологии и синтаксиса. Понимая эту проблему, мы, тем не менее, не пытаемся ее решить, поскольку корпусные проблемы, о которых пойдет речь ниже, связаны с орфографией, которая также отталкивается от среднеевропейского представления о выделении слова и т.д.)

В настоящей работе мы затрагиваем некоторые проблемы, связанные с противопоставлением морфологии и синтаксиса, с которыми нам пришлось встретиться в ходе разработки электронного корпуса адыгейских текстов. Адыгские языки в целом принято характеризовать как полисинтетические [Кумахов 1964] – и действительно, например, адыгейский язык обнаруживает многие специфические черты, распространенные в полисинтетических языках (вроде слабого противопоставления частей речи, нарушения сформулированных на европейском материале законов анафоры и проч.; см. [Lander, Testelets, forthc.]). Адыгейская морфология, подобно синтаксису, в значительной степени допускает значимую перестановку единиц, которыми она оперирует (морфем и сочетаний морфем), а также рекурсию – допустимое построение единицы на основе элементов, включающих единицу того же уровня (ср., например, присоединение бенефактивной морфологии к основе, уже включающей бенефактивный префикс, в *зы-фы-шъу-фэ-с-тхы-рэ-р* [REL.IO-BEN-2PL.IO-BEN-1SG.ERG-писать-DYN-ABS] ‘то, для чего я вам это пишу’). Кроме того, по крайней мере в некоторых фрагментах грамматики адыгейская морфология чрезвычайно продуктивна и, по-видимому, легко допускает построение словоформ в процессе речи.

Адыгейский корпус – электронное собрание адыгейских текстов, разрабатываемое в рамках проекта «Электронная документация полисинтетического языка». Насколько нам известно, это первый опыт открытого корпуса, в котором возможен поиск не только по словам, сочетаниям символов, аффиксам, лексемам или грамматическим характеристикам словоформ, но и по порядку аффиксов и сочетаниям морфем, в том числе разрывным. Для этого в корпусе введены морфологические глоссы – условные обозначения аффиксов (например, PST для прошедшего времени, ABS для абсолютивного падежа) и классов аффиксов (например, APP для аппликативов – префиксов актантной деривации,

CASE для падежных показателей), а также условные обозначения для границ морфем (дефис), любой ненулевой последовательности морфем (звездочка), одной морфемы (вопросительный знак) и т.д.; см. подробнее Arkhangelskiy, Lander 2015; 2016. Такая поисковая система позволяет искать, например словоформы с двумя аппликативными префиксами, разделенными одной морфемой, которые при этом содержат падежный префикс абсолютива (APP-?-APP-*-ABS) и находить примеры вроде *щы-зэ-х-а-ща-гъэ-р* [LOC-REC.IO-LOC-3PL.ERG-вести-PST-ABS] ‘то, что там провели’.

Естественно, поиск по сочетаниям морфем, равно как и морфологический парсер, предлагающий поморфемный анализ, работает только в пределах словоформы. Обычно выделение словоформы не представляет сложностей: существует ряд морфонологических и морфологических правил, которые действуют только в пределах слова (см., например, [Аркадьев и др. 2009]). Так, например, хотя порядок морфем и допускает некоторую вариативность, каждая морфема может быть отнесена к своей морфологической зоне, причем порядок зон внутри словоформы является строгим, определяя ее границы.

В орфографии пробелы в большинстве случаев соответствуют границам словоформ. Соответственно, морфологический анализатор может работать с орфографическими словами, строя гипотезы об их устройстве на основе возможных сочетаний известных ему аффиксов с известными ему корнями, данными в словаре. Тем не менее в адыгейском языке есть и конструкции, в которых это не так и которые мы рассмотрим далее.

2. Наиболее заметной конструкцией такого рода является именной комплекс – сочетание имени и (некоторых) его определений. Как показано в [Lander 2017], канонический именной комплекс обладает свойствами единого слова. Так, модифицирующие его префиксы присоединяются перед всем сочетанием, а модифицирующие его суффиксы – после, несмотря на то, что части именного комплекса в большинстве случаев пишутся раздельно; ср. *т-и-къэлэ кIэракI-эу* [1PL.P-POSS-город=красивый-ADV] ‘как наш красивый город’. Очевидно, что пользователь может нуждаться в поиске, который рассматривал бы именной комплекс как одно слово (например, если его интересуют все случаи, когда посессивная морфология сочетается в одном слове с падежной морфологией в одном слове). С учетом этого, при разработке корпуса можно идти двумя путями.

Во-первых, можно научить анализатор воспринимать предполагаемые именные комплексы как единые слова. Именно этот путь реализован в Адыгейском корпусе на данный момент. При выполнении некоторых требований (неоформленность непоследнего слова сочетания суффиксами, однозначно указывающими на конец слова, и неперв ого слова сочетания префиксами, однозначно указывающими на начало слова, но также и некоторых других) анализатор рассматривает определенные сочетания как именные комплексы и включает их в результаты поиска как единые слова. Проблема в том, что на данный момент это делается в ущерб другим конструкциям: одно и то же сочетание не может одновременно трактоваться в корпусе как единое слово и как сочетание слов. Кроме того, пользователю приходится проделывать ручную работу по отделению реальных именных комплексов от ложных.

Во-вторых, можно дать подробные инструкции пользователю корпуса, как отыскивать именные комплексы (формулируя фактически те же самые ограничения, что использует морфологический анализатор) – таким образом, оставляя пользователю дополнительную ручную работу, но не затрагивая правильность разбора других конструкций. Плюсом этого решения является то, что лексические части именного комплекса, в принципе, могут проявлять и свойства отдельных слов – в частности, они могут сами состоять из нескольких морфологических зон, хотя внутри именного комплекса они входят в единую зону основы.

В-третьих, можно ограничить автоматическое выделение именных комплексов исключительно морфонологическими правилами, хорошо выделяющими слово, но в этом случае затронутыми окажутся лишь часть именных комплексов, в которых эти морфонологические правила реально проявляются. Точность выделения именных комплексов должна повыситься (поскольку, как можно ожидать, ложные именные комплексы включаться в разбор не будут), но многие комплексы останутся «за бортом» – и таким образом, их подбор тоже потребует ручной работы.

Мы не видим идеального выхода из этого положения. Очевидно, что любой выбор предполагает значительное количество ручной работы пользователя. Возможно, наиболее приемлемым является второй вариант, поскольку при нем пользователь по крайней мере сохраняет наибольший контроль над исследованием.

Дополнительным аргументом в пользу второго – «нулевого» – варианта служит унификация с некоторыми глагольными конструкциями. В адыгейском языке имеются сложные конструкции, в которых в качестве вершины выступает некоторый элемент, в той или иной степени грамматикализованный (например, глагол *хъу-* ‘случаться’), а лексический глагол примыкает к нему, но остается неоформленным – не принимает никаких специальных показателей, которые свидетельствовали бы о подчинении; ср. *я-Іэ хъу-гъэ* [3PL.IO+POSS-быть] случаться-PST ‘у них появился’ (см., о подобных конструкциях, например, [Кушнир 2011]). Для поиска в корпусе и морфологической разметки такие конструкции не представляют трудности, поскольку видимого несоответствия расположения аффиксов относительно модифицируемых ими элементов в них не наблюдается. Вместе с тем, можно предположить, что в ряде случаев такие построения фактически образуют единые словоформы. Нам не очевидно, что эти конструкции могут получить единообразную трактовку – и поэтому мы предпочитаем в данном случае слепо следовать орфографическим принципам.

3. В то время, как анализ сочетания орфографических слов поднимает проблему избыточности трактовки именных комплексов как единых грамматических слов, существует и обратная проблема: неанализируемые сочетания основ внутри орфографических слов. Эта проблема тоже связана в первую очередь с именным комплексом.

По правилам адыгейской орфографии односложные определения, входящие в именной комплекс, регулярно не отделяются пробелом от определяемого слова; ср. *кІэла-бэ* [парень-много] ‘много парней’. Некоторые такие сочетания, безусловно, лексикализованы, но то же можно сказать и о многих образующих именной комплекс сочета-

ниях, которые пишутся через пробел. Поскольку модификация в рамках именного комплекса абсолютно продуктивна, на данный момент единственный алгоритм, позволяющий распознать такие сочетания, состоит в полном их исчислении.

Наконец, последнее рассматриваемое нами явление – именная инкорпорация, при которой инкорпорируемое слово присоединяется к глагольной основе [Багирокова, Ландер 2015]. В результате этого образуется слово, имеющее хабитуальное значение и в качестве такового чаще всего используемое как имя – поэтому инкорпорация традиционно описывается как именное словообразование; ср. *бэдзэ-гъэ-лIэ-* [муха-CAUS-умирать] ‘умерщвляющее мух’ = ‘средство против мух’. В отличие от прочих обсуждаемых здесь процессов, именная инкорпорация не является абсолютно продуктивной – многие потенциально допустимые сочетания основ носители оценивают как несуществующие, допустимые в лучшем случае в качестве языковой игры. При этом хотя чаще всего в тексте основы при инкорпорации не разделяются пробелом, носители иногда все же пишут соответствующие образования в два слова. Какой-либо полный список слов, образованных в результате инкорпорации, отсутствует, что также является проблемой для анализатора.

4. Выше мы рассмотрели несколько случаев, когда специфика орфографического слова в адыгейском тексте вызывает сложности при автоматическом морфологическом анализе. Фактически выделяются два класса явлений:

- Последовательность орфографических слов может себя вести как единое слово. Для таких явлений мы считаем, что оптимальным является следование орфографической традиции, поскольку такие случаи вычленяются через задание ограничений при поиске в корпусе.
- Единое орфографическое слово нуждается в разбивке на несколько лексических единиц. На данный момент рабочий анализ для таких построений основан на простом исчислении возможных сочетаний.

Указанные трудности, как кажется, в значительной степени мотивированы типологической спецификой адыгейского языка. В силу того, что адыгейская морфология отчасти берет на себя функции синтаксиса, морфологические единицы получают неожиданно широкие (по сравнению с языками среднеевропейского стандарта) комбинаторные возможности. У аффиксов это проявляется, в частности, в том, что они часто могут присоединяться как к именным, так и к глагольным основам, что приводит к сложностям в противопоставлении имен и глаголов. Однако то же явление наблюдается и при сочетании лексически единиц, традиционно ассоциируемых с отдельными словами. Именно из-за этого и возникает ситуация, когда простое следование делению на орфографические слова может вызывать трудности при морфологическом анализе.

Исследование выполнено при частичной финансовой поддержке РФФИ (проект № 15-06-07434а «Электронная документация полисинтетического языка»).

Литература

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

Багирокова, Ландер 2015 – *И. Г. Багирокова, Ю. А. Ландер*. Именная инкорпорация в адыгейском языке // IV Международный симпозиум лингвистов-кавказоведов. Вопросы структуры основы и корня иберийско-кавказских языков. Тбилиси, 2015. С. 104–106.

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

Кушнир 2011 – *Е. Л. Кушнир*. Грамматикализация вспомогательных глаголов с чистой основой в адыгейском языке // Д.П. Бак, В.В. Минаев (ред.). Полевые исследования студентов РГГУ. Этнология. Фольклористика. Лингвистика. Религиоведение. Вып. VII. М.: Изд. РГГУ, 2011. С. 216–230.

Ландер 2011 – *Ю. А. Ландер*. Подходы к полисинтетизму // Вестник РГГУ. Сер. Филологические науки. Языкознание. № 11(73)/11. (Московский лингвистический журнал. Т. 13.) 2011. С. 102–126.

Arkhangelskiy, Lander 2015 – *T. Arkhangelskiy, Yu. Lander*. Some challenges of the West Circassian polysynthetic corpus // Working papers by NRU HSE. Series WP BRP "Linguistics". No. 37/LNG/2015.

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

de Reuse 2009 – *W. J. de Reuse*. Polysynthesis as a typological feature. An attempt at a characterization from Eskimo and Athabascan perspectives // M.-A. Mahieu, N. Tersis (eds). Variations on Polysynthesis: the Eskaleut Languages. Amsterdam: John Benjamins, 2009. P. 19–34.

Lander 2017 – *Yu. Lander*. Nominal complex in West Circassian: Between morphology and syntax // Studies in Language. Vol. 41, No. 1. 2017. P. 76–98.

Lander, Testelets, forthc. – *Yu. Lander, Ya. Testelets*. Adyghe (Northwest Caucasian) // M. Fortescue et al. (eds) The Oxford Handbook of Polysynthesis. Oxford: Oxford University Press, forthc.