DOI:10.30842/ielcp230690152202

А. В. Андреев

СИНТАКСИЧЕСКИ УПРАВЛЯЕМАЯ РАЗМЕТКА НЕСТАНДАРТНЫХ ТЕКСТОВ (на материале «Катехизиса» 1595 г. М. Даукши)[1]

В статье рассматривается метод бессловарного морфологического анализа, специально приспособленный для текстов на флективных языках с высокой вариативностью, в первую очередь – для разметки старинных текстов. Суть метода заключается в применении синтаксических ограничений к набору возможных морфологических интерпретаций финалей слов. Неоднозначность разбора также уменьшается посредством применения универсальных ограничений, таких как принцип проективности и минимизация набора возможных основ. Описываются результаты применения метода к разметке «Катехизиса» 1595 г. М. Даукши.

Ключевые слова: морфологическая разметка, исторические корпуса, литовский язык, языковая вариативностью.

В настоящее время при составлении корпусов текстов морфологическое аннотирование является практически обязательным этапом. В то время, как для современных стандартизованных языков задача автоматического аннотирования не представляет особой трудности, в других случаях (таких как исторические, диалектные, разговорные корпуса и т. п.) ситуация оказывается не столь определенной. Связано это с тем, что большинство современных методов автоматического морфологического разбора требует наличия формализованной модели морфологии анализируемого языка и, что самое главное, грамматического словаря. В случае древних текстов ни того, ни другого, как правило, не существует; более того, отсутствие жесткой языковой нормы и «несовершенство» орфографии делают создание словаря, аналогичного, допустим, «Грамматическому словарю русского языка» А. А. Зализняка, достаточно нетривиальной задачей. Впрочем, даже и в отношении «стандартных» текстов существует одна важная задача, решение которой далеко от окончательного, а именно – проблема разрешения морфологической омонимии, особенно актуальная для языков «клас-

[1] Работа выполнена при финансовой поддержке Российского научного фонда, грант № 17-18-01624

сического» индоевропейского типа. Таким образом, если речь идет о разработке метода автоматической аннотации текста, то было бы желательно, чтобы механизмы снятия омонимии в нем уже были в том или ином виде изначально заложены.

Сейчас преобладают два основных подхода к автоматическому морфологическому анализу «нестандартного» языкового материала:

- полуавтоматический прецедентный разбор (Arkhangelsky, Mishina, Pichkhadze 2014). При этом анализатор предлагает пользователю выбор из возможных вариантов аннотации, основываясь на словаре, морфологической модели и информации о контекстах данной формы, но окончательное разрешение неоднозначностей, а также обработка нераспознанных или неправильно проанализированных форм остается целиком ответственностью пользователя-лингвиста. Этот подход в принципе позволяет добиться идеальной точности, поскольку и словарь, и модель развиваются по мере анализа текста, однако он требует значительных человеческих ресурсов и, по очевидным причинам, лучше всего работает на достаточно больших корпусах.
- искусственная стандартизация текста, т. е. сведéние его посредством некоторых формальных преобразований к тексту на языке с известной морфологической моделью и словарем, например, старолитовского к современному литовскому (Gelumbeckaitė, Šinkūnas, Zinkevičius 2012) или русского языка XVIII в. к современному русскому (Polyakov 2012). Разумеется, такой подход возможен только в том случае, если лексико-грамматические различия между языками не слишком велики. Достоинством данного метода является, прежде всего, то, что он позволяет получить хоть какую-то морфологическую аннотацию, не затрачивая больших усилий, однако исследователя здесь подстерегает существенная методологическая проблема: в процессе «осовременивания» текста существует большая вероятность устранить, собственно, те грамматические явления, изучение которых составляет основную цель создания исторических корпусов.

Что касается статистических методов морфологического анализа и снятия омонимии, то их применение к историческим корпусам несколько ограничено вследствие малого объема и высокой вариативности таких текстов; впрочем, есть и

успешные примеры использования этих методов, см. обзор в (Gavrilova, Shalganova, Lyashevskaya 2016).

Нами был предложен метод *бессловарной* автоматической морфологической разметки, который опирается на априорную приблизительную модель морфологии и синтаксиса (Andreev 2014). Ниже мы опишем дальнейшее развитие этого метода. Материалом для исследования послужило оцифрованное нами издание «Катехизиса» под ред. Й. Палёниса (DK 1995), с частичной ручной морфологической разметкой и подготовленный в Институте литовского языка цифровой корпус (DK 2006).

Первым этапом обработки текста является частичная орфографическая нормализация на основании недавних исследований орфографии М. Даукши (Andreev 2013; Hock 2014). Следует заметить, что целью такой нормализации является не полное устранение вариативности (как в случае словарных методов), а устранение тех элементов графики, которые, как нам представляется, несущественны для определения грамматического значения слов, что позволяет упростить описание морфологических правил. Таким образом, в принципе этап нормализации является факультативным; в том случае, если окажется, что те или иные графические признаки все-таки релевантны для грамматического описания, они могут быть легко учтены.

Орфографическая нормализация включает в себя следующие правила:

- упорядочиваются чисто графические варианты *u*/*v*/*w*, *i*/*j*, *s*/*ſ*;
- устраняются практически все диакритики над гласными и согласными, так как функция их во многом остается неясной (разумеется, кроме очевидных *č* и *ž*; в этом случае происходит только унификация нескольких вариантов диакритики, а также кроме точки над согласными в конце слова, которая имеет довольно четкую морфологическую значимость, например, *-ti̇* как показатель инфинитива);
- унифицируется также написание *s*/*ss*, *y*/*i*, трех способов записи звука /æ/ (*ia*, e^a, *ę*), назализованных *ą*, *ę*, *į*, *ų* в начале и середине слова;
- вводятся отдельные обозначения для *i* как показателя мягкости предыдущего согласного и для *i* и *u.* Таким образом, нормализованный текст приближается к грубой фонологической транскрипции.

Морфологическая модель опирается на описание грамматики языка Даукши в (Palionis 1995: 50–62; Zinkevičius 1988: 186–190). В первую очередь она состоит из набора стратифицированных правил вида:

Уровень: [Чередование] + Финаль → Набор морфологических тегов / Новый уровень

Набор всех возможных морфологических разборов получается последовательным применением к данной форме всех правил уровня 0 с совпадающей финалью (мы намеренно не употребляем здесь термина суффикс, потому что в правилах могут фигурировать не обязательно элементы, обладающие самостоятельным морфологическим статусом, но просто диагностически значимые сегменты). Финаль отсекается, к получившейся основе применяется правило чередования, если оно есть, и затем результат опять пропускается через набор правил на том же или следующем уровне, и так повторяется до тех пор, пока хотя бы одно правило может быть применено [2]. Полученный набор признаков затем сверяется с таблицей всех возможных морфологических граммем, и разборы, приводящие к невозможным сочетаниям отбрасываются.

Например, правило для возвратного постфикса выглядит следующим образом:

0: [*ie* → *i*; *ů* → *u*] + *s* → REFL / 0

что означает: «у формы, имеющей на конце -*s*, устранить его; затем применить закон Лескина[3]; добавить возвратность к списку грамматических значений; остаться на том же уровне». Последнее необходимо, чтобы можно было применить потом правило для невозвратных окончаний, например:

0: *u* → PERS1, SG, CONJ1 / 1

С другой стороны, например, правила для форманта -*k* повелительного наклонения отличаются на нулевом и первом уровне, потому что -*k* в конце слова без дополнительной финали означает именно 2 лицо ед. числа:

0: *k* → IMP, PERS2, SG / 2

1: *k* → IMP / 2

[2] В настоящее время префиксация в модели не учитывается; это чисто техническое ограничение, которое со временем будет устранено.

[3] По поводу соотношения между синхронными и диахронными правилами см., например, Antilla 1989: 130.

Отдельные служебные слова и неправильные формы (например, многие формы глагола ‘быть’ идентифицируются специальными правилами словарного типа, например:

0:*æst* = *bu*[4] → VERB, PRES, INDIC, PERSON3, COPULA, ATHEM.

Общая таблица граммем включает, помимо традиционных морфологических значений, также некоторые являения, которые обычно рассматриваются на синтаксическом уровне как клитики (Ambrazas 2006: 80, 87–89), в первую очередь, вопросительная частица *-gu* и усилительная частица *-ġ*, которым в нашей модели приписываются соотвестственно признаки interr и emph. Приведем здесь таблицу граммем целиком, отметив, что технически она устроена несколько сложнее, как набор теоретико-множественных выражений, так что граммемы разных классов могут влиять друг на друга (например, в настоящем времени возможны только окончания 1-3 спряжений (признаки CONJ1-3), в прошедшем – только 3 и 4 (признаки CONJ3-4), а в остальных временах – только CONJ2); в таблице это взаимовлияние отражено не полностью.

Отметим несколько ограничений данной модели. Во-первых, граммемы, выражаемые в литовском языке только аналитически (в первую очередь, перфект) не выделяются отдельно (возможно, в дальнейшем это ограничение будет снято). Во-вторых, нашей целью было прежде всего выделение грамматических значений, а не лемматизация, которая вообще вряд ли может быть надежно проведена автоматически в условиях значительной языковой вариативности. Поэтому леммой считается в контексте настоящей работы неразложимая по правилам цепочка графем + классифицирующие признаки (часть речи, тип склонения / спряжения). В силу формальных ограничений на вид чередований некоторые формы, логически относящиеся, безусловно, к одной лексеме будут представлены разными леммами; это касается в первую очередь глагольных форм прошедшего *vs.* настоящего времени.

[4] *bu* используется здесь исключительно как условная метка, объединяющая все формы глагола ‘быть’ в синхронном описании, а не как указание на и.-е. корень $*b^h\bar{u}$, который, по общему мнению, презентного значения не имел.

<table>
<tr><td rowspan="3">Части речи:
NOUN
PRON
ADJ</td><td rowspan="3">Падежи:
NOM
GEN
DAT
ACC
INSTR
LOC
ILL
ADESS
ALL
VOC</td><td rowspan="3">Число:
SG
PL
DUAL</td><td rowspan="3">Род:
M
F</td><td>Тип склонения:
DECL1-6</td><td colspan="3" rowspan="2"></td></tr>
<tr><td rowspan="2">Причастия:
PART. ACT
PART. PASS.
SEMI-PART</td><td>Нескл. формы:
COMPAR
ADV</td></tr>
<tr><td>DET</td><td colspan="2"></td></tr>
<tr><td rowspan="2">VERB</td><td rowspan="2">Лицо:
PERS1
PERS2
PERS3</td><td rowspan="2">Модель управления ~ падеж
COPULA</td><td rowspan="2">Спряжение:
CONJ1-4
ATHEM</td><td>Наклонение:
IND
SUBJ
OPT
IMP</td><td>Время:
PRES
PRET
FUT</td><td rowspan="2">Клитики:
INTERR
EMPH
REFL</td></tr>
<tr><td colspan="2"></td></tr>
<tr><td></td><td>Нелич. формы:
INF
SUP</td><td colspan="4"></td></tr>
<tr><td>PREP</td><td colspan="7"></td></tr>
<tr><td>ADV
CONJ</td><td colspan="7"></td></tr>
</table>

Синтаксическая модель претерпела значительные изменения по сравнению с предложенной нами ранее (Andreev 2014). Изначально мы опирались на подход контекстно-свободных грамматик, однако практика показала, что для наших целей гораздо удобнее использовать грамматики зависимостей, хотя в некоторых случаях они вынуждают прибегать к приемам, не вполне корректным с точки зрения строгой лингвистики. Следует заметить, что нашей целью не является получение автоматической синтаксической разметки в полном объеме, особенно в отношении синтаксиса целого предложения. Очевидно, что во многих случаях такая разметка не может быть получена без обращения к семантике. Нашей главной задачей является использование информации о синтаксических связях в качестве ограничений на возможные грамматические значения отдельных форм, поэтому полученное синтаксическое дерево

практически всегда будет *допустимым* с точки зрения грамматики, но не обязательно тем самым, которое уместно в данном контексте. Математически эта задача решается т. н. методом удовлетворения ограничений в конечных доменах (Triska 2012).

В рамках данной модели рассматривается один недифференцированный вид подчинительной синтаксической связи. Каждая форма имеет ровно одного синтаксического хозяина (за исключением формы-вершины) и произвольное число зависимых слов. Допустимые связи между формами, относящимися к тем или иным граммемам, описываются набором правил (в отличие от морфологических эти правила достаточно универсальны и могут с небольшими изменениями использоваться и для других языков индоевропейского типа. Правила могут быть трех разновидностей:

- обязательная валентность («если форма A обладает набором признаков *X*, то от нее должна зависеть хотя бы одна форма B с набором признаков *Y*»)
- факультативная валентность («если форма A с набором признаков *X* зависит от формы B, то форма B должна обладать одним из наборов признаков Y_1, Y_2, … Y_n
- созависимость («если форма A с набором признаком *X* зависит от того же хозяина, что и B, то форма B должна обладать одним из наборов признаков Y_1, Y_2, … Y_n

Все правила могут быть ограничены позиционно («хозяин слева/справа от зависимого слова», «хозяин непосредственно примыкает к зависимому слову»). Особенностью «Катехизиса» является вопросно-ответная структура текста, причем часто ответ является эллиптическим предложением по отношению к вопросу. Поэтому в нашей модели предусмотрено проведение синтаксических связей через границу предложения (это тоже представляется как позиционное ограничение особого рода).

Например, правила могут выглядеть следующем образом:

- PREP_X *requires rightward* NOUN, X, X ∈ {GEN, ACC, DAT, INSTR}, т. е. «за предлогом, требующим падежа X должно обязательно следовать существительное в форме этого падежа»
- VERB, GOVERN_X *requires following* X, X ∈ {NOM, GEN, ACC, DAT, INSTR, LOC, ILL}, т. е. «если глагол управляет каким-

либо падежом, то форма в этом падеже должна появиться в этом или следующем предложении»

- $\text{PERSON}_{1,2,3}$ *allows* NOM, т. е. «личная форма глагола допускает при себе зависимое слово в им. падеже в этом же предложении (подлежащее)»

Вершиной синтаксического дерева всегда предполагается глагол в личной форме, возможно, в другом предложении. Это делает разбор некоторых предложений невозможным, однако в нашем материале назывные предложение – это в основном разного рода заголовки, так что на практике это ограничение не очень существенно. Особого внимания требуют случаи, в которых вид и направление синтаксической связи не вполне однозначно определяются. Так, частицы всегда считаются синтаксически примыкающими к предыдущему слову, что для литовского языка практически всегда дает лингвистически осмысленный результат. Слово из эллиптического ответа связывается с вопросительным словом из предложения-вопроса, что не вполне традиционно, но кажется разумным решением. В случае сочинительных конструкций нам пришлось прибегнуть к техническому приему и предположить, что сочинительный союз наследует признаки соединяемых слов. Иными словами, например, для союза *ir* существует набор правил, требующих слева и справа от него слов с совпадающими наборами признаков:

CONJ(NOUN, X) *requires rightward* NOUN, X & *leftward* NOUN, X, X ∈ {NOM, GEN, ACC, DAT, INSTR...}

CONJ(ADJ, X, Y) *requires rightward* ADJ, X, Y & *leftward* ADJ X, Y, X ∈ {NOM, GEN, ACC, DAT, INSTR...}, Y ∈ {M, F}

CONJ(VERB, X) *requires rightward* VERB, X & *leftward* VERB, X, X ∈ {PERS1, PERS2, PERS3, INF}

После того, как возможные синтаксические деревья построены, лишние варианты убираются посредством применения дополнительных универсальных ограничений. Во-первых, мы используем принцип проективности, требующий, чтобы стрелки синтаксических зависимостей не пересекались (этот принцип, как известно, не всегда применим, но в текстах простой структуры отклонения от него чрезвычайно редки). Во-вторых, предполагается, что в одном предложении не могут встретиться два омонима, т. е. если какая-то форма встречается в предложении дважды, то она оба раза должна принадлежать к одной

лемме. В общем случае это, разумеется, неверно, но практически, особенно при отсутствии в тексте элементов языковой игры, применение такого ограничения позволяет значительно уменьшить число возможных разборов в случаях типа [*Jesus*] *ira tikras diewas ir tikras žmogus* ‘[Иисус] истинно бог и истинно человек’. Наконец, последнее правило, применяемое глобально ко всему анализируемому тексту, состоит в том, чтобы всегда предпочитать разборы, которые дают суммарно наименьшее количество лемм.

В завершение рассмотрим пример работы описанной процедуры:

M. *Kaiṕ iżpíldiſſime*^{a} *ke*^{a}*twîrtą priſsâkimą / apę miłêiimą téwo ír môtinos?*

Mo. *Túrime*^{a} *ių kłausíṫ / turéṫ iůs page*^{a}*rbime*^{a} */ tarnáuṫ iiemus / ir paßełpṫ iůs.*[DK 81.11–17, написание несколько упрощено по техническим причинам]

‘Как исполним четвертую заповедь о любви отца и матери? Должны их слушать, иметь их в почести, служить им и помогать им [в лит. – acc.]’

После применения процедур нормализации орфографии получаем:

kaĭp iżpildisimæ kætwirtą prisakimą apę mileĭimą tewo ir motinos?

turimæ ĭų klaŭsit^{j} *turet*^{j} *ĭůs pagærbimæ tarnaŭt*^{j} *iĭemus ir paßelpt*^{j} *ĭůs*

Применяя морфологические правила приходим к следующим результатам:

kaĭp (ADV, INTERR) *iżpildisimæ* (VERB, FUT, PERSON3, PL | NOUN, LOC, SG, DECL1) *kætwirtą* (ADJ|NOUN, ACC, SG, DECL1|DECL2) *prisakimą* (ADJ|NOUN, ACC, SG, DECL1|DECL2) *apę* (PREP(ACC)) *mileĭimą* (ADJ|NOUN, ACC, SG, DECL1|DECL2) *tewo* (NOUN, GEN, SG, DECL1 | VERB, PRES|PRET, CONJ3) *ir* (CONJ) *motinos* (NOUN GEN, SG | NOM, PL, DECL2 | VERB. PRES|PRET, CONJ3, REFL) *turimæ* (VERB, PRES, CONJ2, PERS1, PL | NOUN, LOC, SG, DECL1) *ĭų* (PRON, GEN,. PL) *klaŭsit*^{j} (VERB, INF) *turet*^{j} (VERB, INF) *ĭůs* (PRON, ACC, PL) *pagærbimæ* (VERB, PRES, PERS1, PL, CONJ2 | NOUN, LOC, SG, DECL1) *tarnaŭt*^{j} (INF.) *iĭemus* (PRON, DAT, PL) *ir* (CONJ) *paßelpt*^{j} (INF) *ĭůs* (PRON, DAT, PL).

Общее количество возможных разборов составляет, следовательно, 16384. Однако, к счастью, мы имеем здесь

много однозначно идентифицируемых слов (предлоги, союзы и инфинитивы). Это сразу же приводит к анализу *mileĭimą* как существительного, следующего за предлогом. После этого необходимость иметь единственный корень синтаксического дерева и требование, чтобы союз *ir* соединял однородные формы сразу дает анализ *tewo* и *motinos* как существительных. По тем же правилам *ižpildisimæ* оказывается глаголом. Таким образом, в этом предложении неоднозначным остается только выбор между существительным и прилагательным для *kætwirtą* и *prisakimą* и их тип склонения. Однако эта неоднозначность снимается минимизацией числа лемм, поскольку *prisakimas* ‘заповедь’ – очень частотное слово в «Катехизисе», так что его частеречная принадлежность и тип склонения определяются другими контекстами. Во втором предложении, так же руководствуясь правилом о союзе *ir*, идентифицируются однородные группы *tarnaĭtʲ iĭemus* и *paßelptʲ ĭůs*. После этого по остаточному принципу также выделяются группы *ĭų klaĭsitʲ* и *turetʲ ĭůs*, а также синтаксический корень: форма *turimæ*. Однако мы не можем сделать выбор между *pagerbimæ* VERB и *pagerbimæ* NOUN, поскольку в других местах текста от этого корня встречается и глагольные, и именные формы. Следовательно, первое предложение в конечном результате получает единственный разбор [[*kaĭp* (ADV, INTERR) *ižpildisimæ* (VERB, FUT, PERSON3, PL)] [*kætwirtą* (ADJ, ACC, SG, DECL1) *prisakimą*]] (NOUN, ACC, SG, DECL1) [*apę* (PREP(ACC)) [*mileĭimą* (NOUN, ACC, SG, DECL1) [*tewo* (NOUN, GEN, SG, DECL1) *ir* (CONJ) *motinos* (NOUN GEN, SG, DECL2)]]], а второе — два, корректный: *turimæ* (VERB, PRES, CONJ2, PERS1, PL) [*ĭų* (PRON, GEN,. PL) *klaĭsitʲ* (VERB, INF)] [*turetʲ* (VERB, INF) *ĭůs* (PRON, ACC, PL) *pagærbimæ* (NOUN, LOC, SG, DECL1)] [[*tarnaĭtʲ* (INF.) *iĭemus* (PRON, DAT, PL)] *ir* (CONJ) [*paßelptʲ* (INF) *ĭůs* (PRON, DAT, PL)]]] и паразитический: [*turimæ* (VERB, PRES, CONJ2, PERS1, PL) <…>] [**pagærbimæ* (VERB, PRES, PERS1, PL, CONJ2) [[*tarnaĭtʲ* (INF.) *iĭemus* (PRON, DAT, PL)] *ir* (CONJ) [*paßelptʲ* (INF) *ĭůs* (PRON, DAT, PL)]]. Отметим, что если добавить специальное морфологическое правило о моделях управления частотного *turét* ‘иметь’, то разбор станет однозначным.

Таким образом, мы можем видеть, что бессловарный морфологический анализ может давать результаты, не уступающими более распространенным словарно-ориентированным методам.

Литература

Ambrazas, V. 2006: Lietuvių kalbos istorine sintaksė. Vilnius: Lietuvių kalbos institutas.

Andreev A. V. 2014: [On a method of automatic morphosyntactic annotation of old texts], In: *Pismenoto nasledstvo i informacionnite tekhnologii. «El'Manuscript-2014». Materiali ot V mezhdunarodna nauchna konferentsiya* [*Proceedings of the 5th International Conference Textual Heritage and Information Technologies, «El'Manuscript-2014»*], Sophia, Izhevsk, 99–101.

Андреев А. В. 2014: Об одном методе автоматической грамматической разметки старопечатных текстов. В сб.: Баранов, В. А., Желязкова, В., Лаврентьев, А. М. (отг. ред.). *Писменото наследство и информационните технологии. «El'Manuscript–2014». Материали от V международна научна конференция*. София, Ижевск, 99–101.

Andreev, A. V. 2013: [Diacritic marks and orthography in M. Daukša's Catechism of 1595: a quantitative aspect], the analysis of the names]. *Indoevropeiskoe yazykoznanie i klassicheskaya filologiya* [*Indo-European linguistics and classical philology*], 17, 16–25.

Андреев А. В. Знаки ударения в орфографии катехизиса М. Даукши 1595 г.: опыт количественного исследования. *Индоевропейское языкознание и классическая филология* 17, 16–25.

Antilla, R. 1989: *Historical and comparative linguistics*. Amsterdam, Philadelphia: John Benjamins Publishing Company.

Arkhangelsky, T. A., Mishina, E. A., Pichkhadze, A. A. 2014: [A system for digital morphological tagging for Old Russian and Church Slavonic texts]. *Palaeobulgarica*, 38 (4), 21–37.

Архангельский, Т .А., Мишина, Е. А., Пичхадзе, А. А. 2014: Система электронной грамматической разметки древнерусских и церковнославянских текстов. *Palaeobulgarica / Старобългаристика*, 38 (4), 21–37.

DK 1995: *Mikalojaus Daukšos 1595 katekizmas*. Red. J. Palionis. Vilnius: Moklso ir enciklopedijų leidykla.

DK 2006: Mikalojus Daukša. *Katekizmas,* 1595. Prepared by M. Šinkūnas, sponsored by Lithuanian State Science and Studies Foundation, 2006. (http://seniejirastai.lki.lt/db.php?source=1).

Gavrilova, T. S., Shalganova, T. A., Lyashevskaya, O. N. 2016: [Lexico-grammatical annotation of the middle russian corpus 1400–1700: a computational approach]. *Vestnik Pravoslavnogo Svyato-Tikhonovskogo gumanitarnogo universiteta. Seriya 3: Filologiya* [*Bulletin of the Orthodox St. Tikhon University of Humanities*], 47 (2) 7–25.

Гаврилова, Т. С., Шалганова, Т. А., Ляшевская, О. Н. 2016: К задаче автоматической лексико-грамматической разметки старорусского корпуса XV–XVII вв. *Вестник Православного Свято-Тихоновского гуманитарного университета. Серия 3: Филология.* 47 (2) 7–25.

Gelumbeckaitė, J., Šinkūnas, M., Zinkevičius,V. 2012: Old Lithuanian Reference Corpus (SLIEKKAS) and Automated Grammatical Annotation. *Journal for Language Technology and Computational Linguistics*, 27 (2). 83–96.

Hock, W. 2014: Untersuchungen zu Daukšas Postille – II. Doppelakzentuierungen, *Archivum Lithuanicum*, 16, 173–202.

Palionis, J. 1995: *Lietuvių rašomosios kalbos istorija*. Vilnius: Mokso ir enciklopediju leidykla.

Polyakov, A. E. 2012: [A stemmer for the pre-reform Russian orthography] In: *Informacionnye tekhnologii i pis'mennoe nasledie: materialy IV mezhdunarodnoy nauchnoy konferentsii (Petrozavodsk, 3–8 sentyabrya 2012 g.)* [*Proceedings of the international conference Information Technologies and Textual Heritage El'Manuscript-12*], Petrozavodsk, Izhevsk, 211–215.

Поляков, А. Е. 2012: Лемматизатор для дореформенной русской орфографии. В сб.: Баранов, В. А., Варфоломеев, А. Г. (отв. ред.) *Информационные технологии и письменное наследие: материалы IV международной научной конференции (Петрозаводск, 3–8 сентября 2012 г.).* Петрозаводск, Ижевск, 211–215.

Triska, M. 2012: The Finite Domain Constraint Solver of SWI-Prolog. In: Schrijvers T., Thiemann P. (eds) *Functional and Logic Programming. FLOPS 2012. Lecture Notes in Computer Science, vol 7294*. Berlin, Heidelberg: Springer, 307–316.

Zinkevičius, Z. 1988: *Lietuvių kalbos istorija.* T. 3. Vilnius: Moklso ir enciklopedijų leidykla.

A. V. Andreev. A method of syntactically-constrained morphological annotation (as applied to «Katechismas» of 1595 by M. Daukša)

In the article, the existing methods of morphological annotation for historical corpora are analysed. A new method of an unsupervised dictionary-free morphological tagging is proposed which is based on applying syntactical dependency constraints to a set of possible morphological interpretations of word finals. The procedure starts with a draft set of orthographical, morphological and syntactic rules that are adjusted and refined as the analysed text is processed. The method is specifically tailored to the highly-inflectionate languages of ‘classical’ Indo-European type. The ambiguity of annotation is further reduced by applying a set of language-neutral constraints, such as the well-known principle of projectivity or the minimization of possible word stems. The application of the method to tagging M. Daukša's Cathechism of 1595 is described.

Key words: morphological annotation, historical corpora, Lithuanian language, linguistic variation.