МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

© Кирило-Методиевски научен център — БАН, 2014
© Ижевский государственный технический университет им. М. Т. Калашникова, 2014
© Авторски колектив, 2014
© Лилия Тошкова — графичен дизайн на корицата, 2014

ISBN 978–954–9787–25–2

Об одном методе автоматической грамматической разметки старопечатных текстов

А. В. Андреев

Автоматическая разметка, морфологический анализ, многовариантный синтаксический анализ, старопечатные тексты, литовский язык

On One Method for Automatic Morphosyntactic Annotation of Early Texts

Artem Andreev

A method is proposed for unsupervised morphosyntactic markup of old texts for which no exact grammar nor vocabulary may be known. The method employs building all possible mappings from text forms into grammemes and then reducing them using a loose context-free (CF) grammar. The forms are further lemmatized based on minimization of morphologic variation. The method has been tested on two old Lithuanian documents from the late 16th century by M. Dauksha and has proven to be rather efficient and accurate (up to 80 %).

Автоматическая грамматическая (морфологическая или синтаксическая) разметка старинных текстов, как правило, наталкивается на ряд трудностей: высокая орфографическая вариативность текста, отсутствие заранее заданного словаря, а зачастую — и отсутствие достаточно точного грамматического описания. Кроме того, такие тексты часто относительно невелики по объему, что затрудняет использование статистических методов анализа текста. Для индоевропейских языков, в особенности балто-славянского типа, проблема усугубляется наличием сложных нелинейных морфологических правил (например, чередований основы).

В докладе предлагается метод автоматического морфосинтаксического аннотирования текста, основанный на использовании частичной грамматической информации. Входными данными для алгоритма разметки служат, помимо собственно размечаемого текста, те сведения, которые обычно известны лингвисту:

- сведения об основных орфографических вариантах,
- инвентарь граммем и приблизительный инвентарь морфов, выражающих эти граммемы,
- сведения о грамматически значимых чередованиях,
- контекстно-свободная грамматика, приблизительно описывающая языки данного типа (отражающая в случае языков балто-славянского типа такие их свойства, как свободный порядок слов, согласование существительных и прилагательных, невозможность вин. падежа в именной группе, инфинитивные и причастные конструкции и т. п.).

© А. В. Андреев, 2014

Не предполагаются известными, таким образом, ни словарь, ни распределение форм по словоизменительным типам, ни информация о синтаксической сочетаемости.

Работа алгоритма состоит из двух фаз. В первой происходит построение вариантов морфосинтаксической аннотации. Для каждой реальной формы в тексте на основе информации об орфографических вариантах и морфах генерируются все теоретически возможные граммемы, которые может иметь данная форма. Затем для каждого предложения происходит редукция вариантов с использованием многовариантного синтаксического анализа [Фитиалов 1998]. Приемлемыми, очевидно, признаются только такие варианты, которые удовлетворяют КС-грамматике. Во второй фазе алгоритма происходит лемматизация аннотированных форм. При этом предпочтение отдается такому распределению форм по лексемам, при котором вариативность морфов и грамматических чередований, выражающих одни и те же граммемы, оказывается наименьшей [Сухотин 1976]. После этого происходит дополнительная редукция вариантов разбора, уже для всего текста, так как некоторые такие варианты могут опираться на отброшенные в фазе два морфологические разборы.

Описанный алгоритм был использован для аннотации первых памятников литовского языка: Катехизиса 1595 г. и Постиллы 1599 г. М. Даукши, на основе грамматических сведений, приведенных в [Palionis 1995; Zinkevičius 1988]. Применение метода показало его достаточно высокую точность и эффективность при условии, что анализируемые тексты состоят, по большей части, из относительно коротких предложений не слишком сложной структуры (что обычно выполняется для текстов учебно-религиозного характера). При наличии только такой грамматической информации, которая была описана выше, точность морфосинтаксической разметки оказывается на уровне 80 %. Наибольшую проблему представляют морфы, текстуально являющиеся частью других морфов, в частности, нулевые окончания (к которым в данном контексте приравнивается и отсутствие морфологических показателей у служебных слов). Точность разметки может быть поэтому существенно повышена до уровня 95 % за счет (а) списка служебных слов, в первую очередь — предлогов, для которых к тому же обычно априорно известна их модель управления, (б) группировкой морфов в парадигматические ряды и дополнительным требованием в фазе 2, чтобы все морфы, сочетающиеся с данной леммой, относились бы к возможно меньшему числу парадигм.

Описанный метод может быть, как можно надеяться, успешно применен для анализа других памятников литовского языка, а также других балтийских и славянских языков (возможность применения его для языков других типов пока остается под вопросом).

Литература

Сухотин 1976 — *Сухотин Б. В.* Оптимизационные методы исследования языка. М.: Наука, 1976.

Фитиалов 1998 — *Фитиалов С. Я.* Алгоритмы многовариантного синтаксического анализа // Структурная и прикладная лингвистика. Вып. 5 / под ред. А. С Герда. СПб.: Изд-во С.-Петербург. ун-та, 1998. С. 50–60.

Palionis 1995 — *Palionis J.* 1595 metų katekizmas ir jo leidimai // Mikalojaus Daukšos 1595 katekizmas. Vilnius: Moklso ir enciklopedijų leidykla, 1995. P. 15–40.

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