ДАНУТА РОШКО, РОМАН РОШКО

Институт славистики Польской академии наук
danuta.roszko@ispan.waw.pl, roman.roszko@ispan.waw.pl

# Польско-литовский корпус

В начале доклада авторы описывают экспериментальные многоязычные корпусы: (а) параллельный и сопоставительный болгарско-польско-литовский корпус и (б) параллельный польско-литовский корпус. Объем первого корпуса (BG-PL-LT) составляет около 2 000 000 словоформ. Работы над этим корпусом уже окончены. Объем второго же корпуса составляет свыше 8 000 000 словоформ. Работы над этим корпусом продолжаются. В качестве текстового материала для включения в этот корпус (PL-LT) были отобраны разнообразные переводы международной художественной литературы (30%), некоторые документы Евросоюза (5%), все и всякие переводы с польского языка на литовский и наоборот (напр.: договоры, нормативные документы, деловые письма, переписка, научные статьи, бланки, каталоги, стенограммы, постановления, судебные определения, решения и др.) (65%).

Во второй части доклада авторы подробно описывают новый, создаваемый в рамках CLARIN-PL, параллельный польско-литовский корпус. Достоинством этого корпуса – по сравнению с выше описанными – является то, что он будет опубликован в сети в 2016 году. В этот корпус будут включены отрывки переводов международной художественной литературы, в том числе также польской и литовской, официальные деловые письма, польско-литовские договоры и документы Евросоюза. Объем этого корпуса достигнет 6 400 000 словоформ.

В последней части доклада авторы описывают новый тип семантической маркировки, который в характере эксперимента включен в этот корпус. Семантическая маркировка, связанная с кванторным описанием значений на уровне предложения, является совсем новым качеством в параллельных корпусах. Применение этой маркировки в многоязычных корпусах приведет к значительному развитию качества машинного перевода.

Пример семантической маркировки:

*Ar tu kada nors kalbėjai ką nors apie didįjį cezarį?*
*Czy kiedykolwiek mówiłeś coś o wielkim Cezarze?*
Ar
<type»(ιx)P(x)» form»tu»\>
<type”(∀$X_1$)P($X_1$)_(state_2)” form”kada nors kalbėjai”\>
<type”(∀$x_1$)P($x_1$) type”ką nors”\>
<type”(ιx)P(x)” form”apie didįjį cezarį”\>
?
Czy
<type”?(∀$X_1$)P($X_1$)_(state_2)” form”kiedykolwiek mówiłeś”\>
<type”?(∀$x_1$)P($x_1$) type”coś”\>
<type”(ιx)P(x)” form”o wielkim Cezarze”\>
?

Как можно заметить, в литовском варианте предложения все кванторные значения однозначны. В польском языке дважды появляется знак «?», который обозначает многозначную вне контекста и ситуации

языковую форму. Значение таких форм преимущественно раскрывается в контексте, но не всегда. В таких случаях мы говорим про кванторную недоговоренность. Например, польская форма *coś* в зависимости от контекста и ситуации может быть средством выражения (свойственной) экзистенциальности (∃x)P(x) (*Coś* ci kupiłem. ‘Я тебе *что-то* купил.’) или (ограниченной) общности ($\forall x_1$)P($x_1$) (*Coś* ci kupię. ‘Я тебе *что-нибудь* куплю.’). Литовские эквиваленты польского *coś* однозначны: *Coś* ci kupiłem. — Aš *kažką* tau nupirkau. (... pūski, pūski kuo stipriau, aš kažką tau nupirkau.) *Coś* ci kupię. — Aš *ką nors* tau nupirksiu.

В семантической маркировке мы различаем для (а) предметных и (б) непредметных переменных следующие значения: [1] единственности, [2-3] два типа экзистенциальности (ограниченной и свойственной) и [4-5] два типа общности (ограниченной и неограниченной).У каждого из выше указанных значений два варианта: однозначный или многозначный (связан с кванторной недоговоренностью). Добавочно для непредметных переменных мы включаем значения: (а) состояния, (б) секвенции/цепочки состояний и событий оконченной состоянием, (в) события, (г) секвенции/цепочки состояний и событий оконченной событием.