**4 ноября**

**Ольга Ляшевская. Корпуса Universal Dependencies, или зачем нам еще один формат синтаксической разметки?**

Корпуса с разметкой частей речи и грамматических признаков давно стали стандартом для проведения лингвистических исследований. Вместе с тем, многие лингвисты-теоретики имеют лишь смутное представление о том, что получается в результате синтаксического парсинга, боятся сложной системы тегов и не понимают, какие данные можно извлечь из таких корпусов (трибанков). К тому же, считают они, трибанки – маленькие и существуют лишь для очень небольшого числа «больших» языков.

Это и правда было так – до недавнего времени. Однако буквально на наших глазах в истории корпусной лингвистики случился поворот: во-первых, были созданы алгоритмы, позволяющие автоматически размечать корпуса в формате грамматики зависимостей с качеством порядка 90%, а во-вторых, возникло движение Universal Dependencies (UD), в рамках которого разрабатываются принципы «квазиуниверсального» формата разметки для разноструктурных языков. Уже сейчас выложены в свободный доступ трибанки для более чем 60 языков.

Между тем, пользователю, привычному к разметке в стиле модели «Смысл-Текст» (см., например, Синтаксический корпус НКРЯ), нужно быть готовым к тому, что разметка UD покажется необычной. Во время доклада мы обсудим некоторые ключевые принципы разметки UD для русского языка и коснемся некоторых сложных и нерешенных проблем разметки для языков разного типа, включая устные и жестовые языки.