**Выделение двуязычных терминов и терминологических словосочетаний на основе параллельного англо-русского корпуса по языкам программирования**

***Кощеева С.С.***

Санкт-Петербургский государственный университет (Санкт-Петербург), Россия

[swetik-1995@mail.ru](mailto:swetik-1995@mail.ru)

**Тезисы**

Параллельные корпусы текстов широко используются в корпусной лингвистике для выделения терминов и терминологических словосочетаний. Современные исследования показывают, что извлечённые из параллельных корпусов двуязычные термины (однословные и многословные) можно применять в качестве основы для машинного перевода, создания терминологических словарей и тезаурусов предметной области, а также для верификации уже зафиксированных значений терминов и оценки адекватности их перевода.

**Цель** данного исследования заключается в том, чтобы на основе параллельного англо-русского корпуса текстов по языкам программирования изучить и сравнить различные методы выделения терминов и терминологических сочетаний.

В корпусной лингвистике были разработаны различные методы извлечения терминов и терминологических сочетаний. Среди методов выделения однословных терминов различают методы ключевых слов и использование моделей дистрибутивной семантики. Основные подходы к извлечению терминологических словосочетаний включают в себя *статистические* (меры ассоциации), *лингвистические* (морфосинтаксические шаблоны) и *смешанные* (сочетание статистических и лингвистического методов).

В ходе экспериментальной части данной работы был создан параллельный англо-русский корпус текстов по языкам программирования: на его основе определялась эффективность статистического и лингвистического (использование морфосинтаксических шаблонов) методов извлечения терминов. Анализ статистического и лингвистического методов выделения двуязычных терминологических биграмм показал, что ни один из методов не отличаются высокими показателями точности. Однако использование морфосинтаксических шаблонов позволяет выделять более разнообразные типы терминологических словосочетаний, в отличие от статистического метода, который выдаёт очень ограниченное количество результатов.

Таким образом, статистический и лингвистический методы не отличаются высокой точностью, но тем не менее могут быть использованы для выделения терминологических словосочетаний. Предполагается, что совместное использование статистического и лингвистического методов (смешанный метод) может значительно повысить качество выделения двуязычных терминов и терминологических словосочетаний в параллельных корпусах текстов.