# Онтологический инжиниринг

## в системах извлечения знаний из текста

*опыт машинного анализа сообщений блога “Живой Журнал” (www.livejournal.ru)*

*Александр Ермаков, ООО «ЭР СИ О»*
ermakov@rco.ruermakov@rco.ru, www.rco.ru

# Знания в текстах: откуда, что и зачем извлекать?

Извлечение из Интернета первичных элементов знания:

а) утверждения (*лекарство Антипилин – полная ерунда*; *вероятная причина свиста под капотом автомобиля в сырую погоду – слабое натяжение ремня генератора*);

б) факта (*после принятия Антипилина может подниматься давление*; *летом 2006 фирма Пежо отозвала 20000 автомобилей из-за возможного возгорания в системе электроусилителя руля*).

Порождение сложного знания из элементов знания:

а) логический вывод, например: *продукт X некачественный* (утверждение), *X - продукт компании Y в 1997* (факт), *Z - технический директор компании Y с 1996 по 1998 годы* (факт), следовательно, *Z - плохой руководитель* (знание);

б) обобщение, например, порождение выводов: *препарат Антипилин имеет меньше побочных эффектов, чем Глипирон* (на основании статистики отзывов больных) или *Типичная причина поломок автомобиля Форд Фокус – засорение бензонасоса* (на основании статистики сообщений автомобилистов).

# Социальные сети в Интернет: главный источник знаний

Блог “**Живой Журнал**” (*http://www.livejournal.ru/*) – сеть электронных дневников пользователей, которые делают записи (посты) в своих дневниках и комментарии на записи других пользователей в своих и чужих дневниках.

По состоянию на лето 2007 года русскоязычная часть блога содержит:

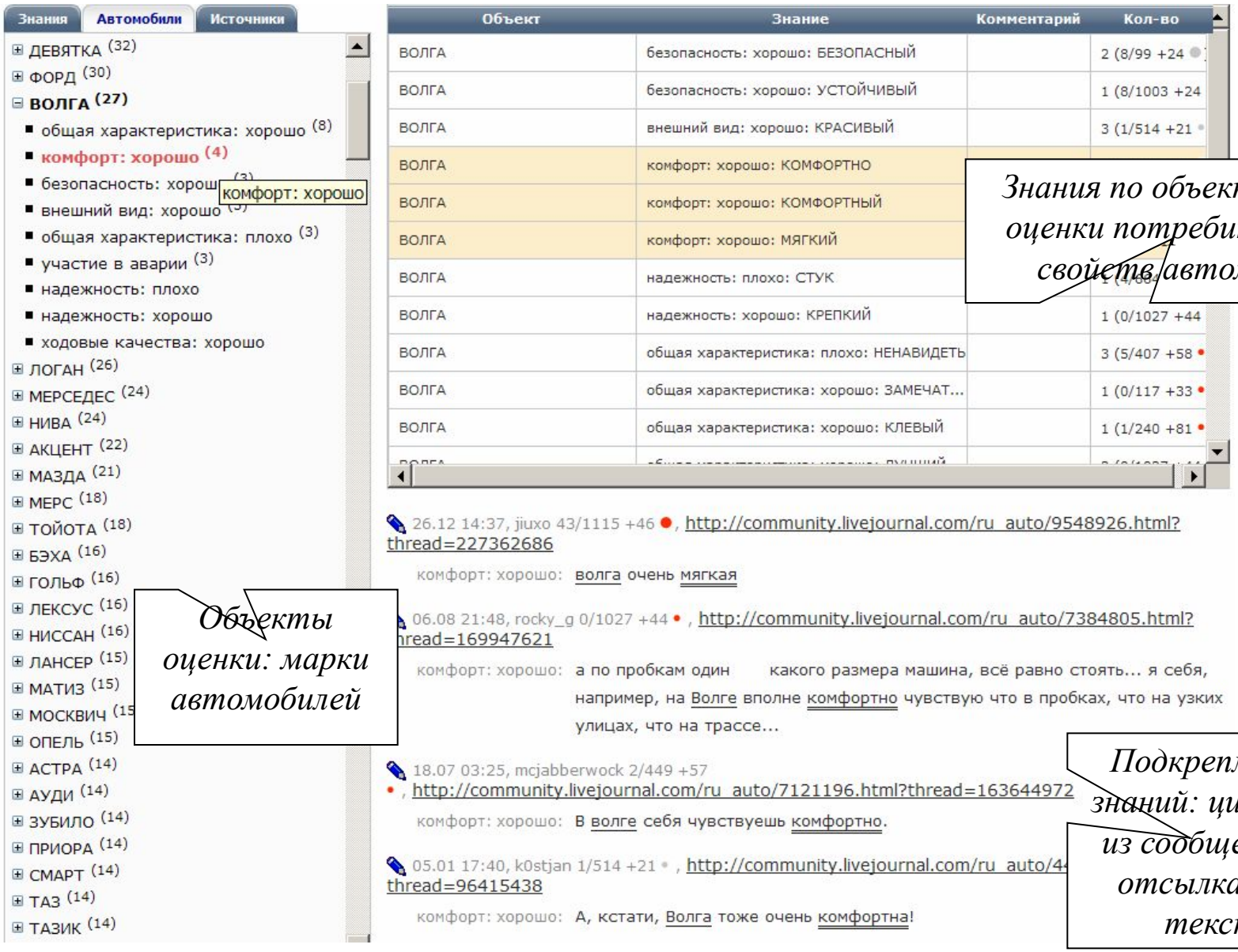
- более 75 тысяч тематических сообществ;
- более 1 миллиона 200 тысяч пользователей;
- в день добавляется около 100 тысяч постов и 400 тысяч комментариев.

Сообщество **auto_ru** (“Все об автомобилях”) – крупнейшее из автомобильных.

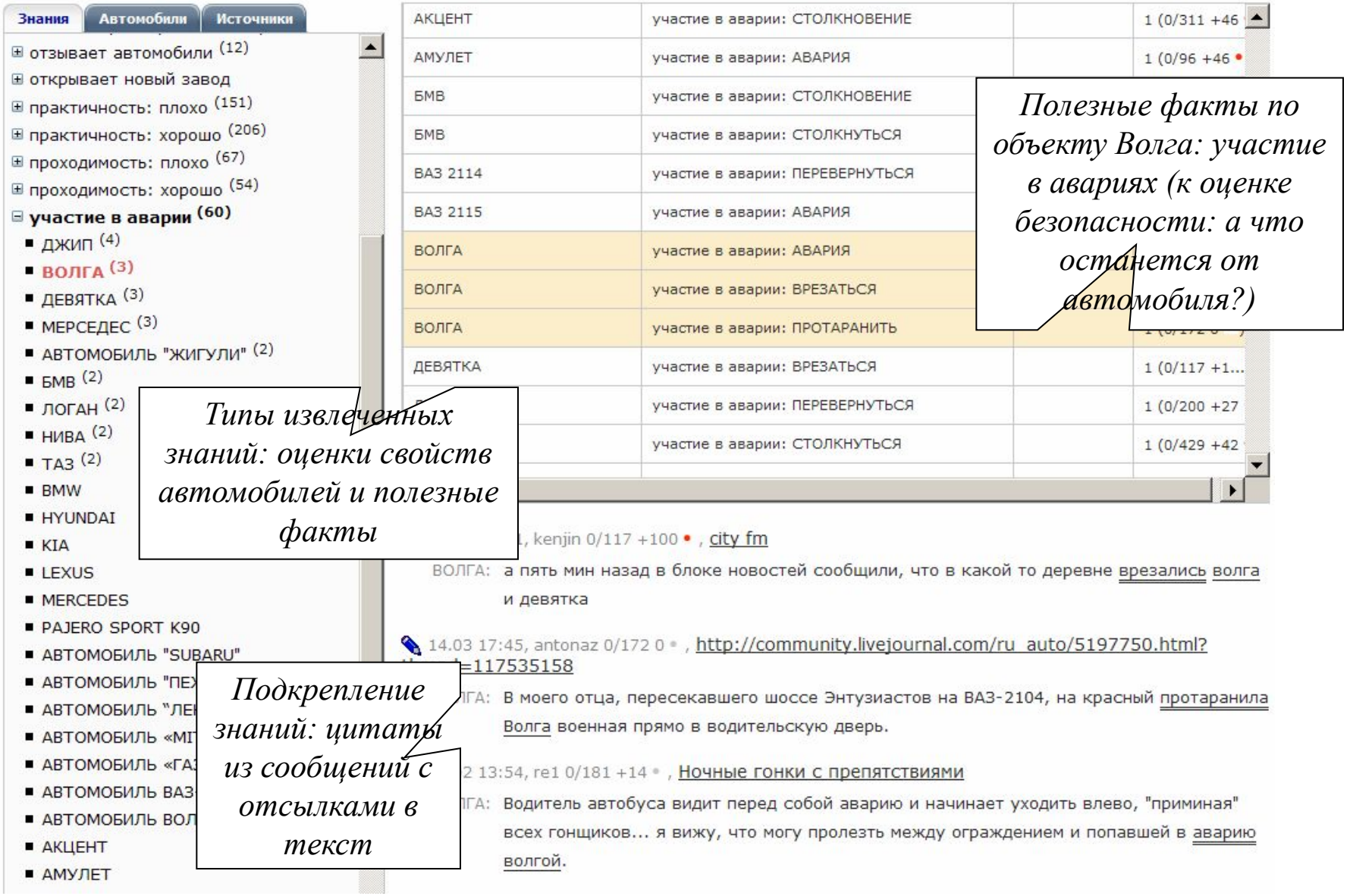
В целом за 2007 год:

- 500 тысяч сообщений, порожденных 19 тысячами постов;
- 3000 авторов постов и 6 тысяч авторов комментариев;
- объем русскоязычного текста около 60 Мбайт.

# Знания об автомобилях из Интернет-сообщества (скриншот 1)



# Знания об автомобилях из Интернет-сообщества (скриншот 2)



# Извлечение знаний из Интернета: оценка потребительских свойств товаров на основании анализа отзывов

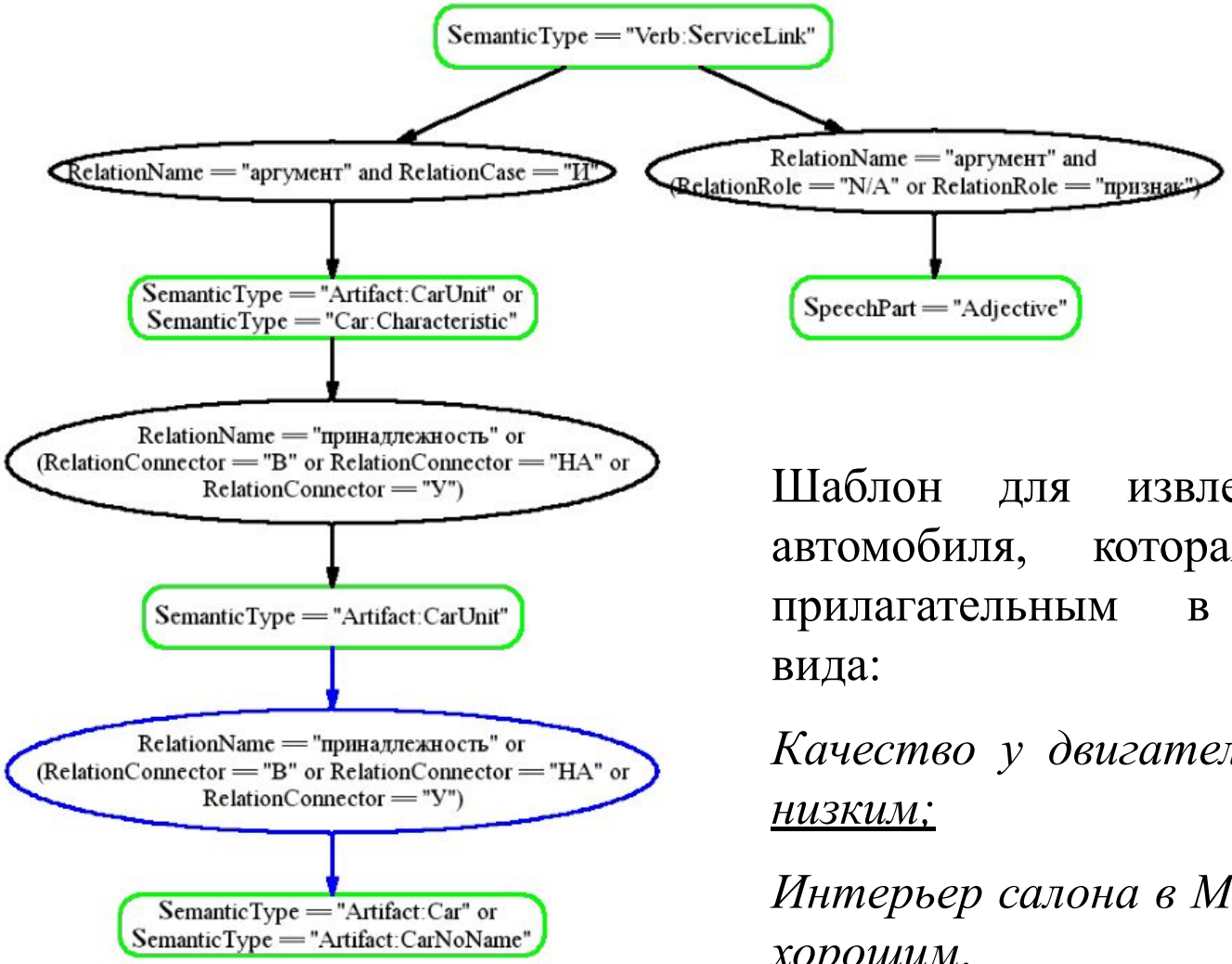
Задача: для каждой модели автомобиля "выловить" положительные и отрицательные отзывы и классифицировать их: за что хвалят/ругают?.

Экспериментальная онтология для оценки автомобилей с точки зрения характеристик (*положительная/отрицательная*) их потребительских свойств. Содержит более 1200 терминов (24 группы):

- 211 наименований узлов (*движок*, *коробка передач, ходовая часть*);
- 71 наименование свойств классифицированны на 8 оцениваемых групп (*ходовые качества*, *комфорт*, *безопасность*, *надежность*, ...);
- 882 наименования оценок характеристик узлов и свойств, включающие прилагательные, существительные, глаголы и наречия (*крутой*, *поломка*, *глючить*, *отстойно*);
- 37 эмоциональных характеристик (*любить*, *жалоба*, *плеваться*).

Синтаксические связи в предложении между 24 группами терминов из онтологии описываются около 100 семантических шаблонов.

# Извлечение знаний: семантическая интерпретация текста (1)

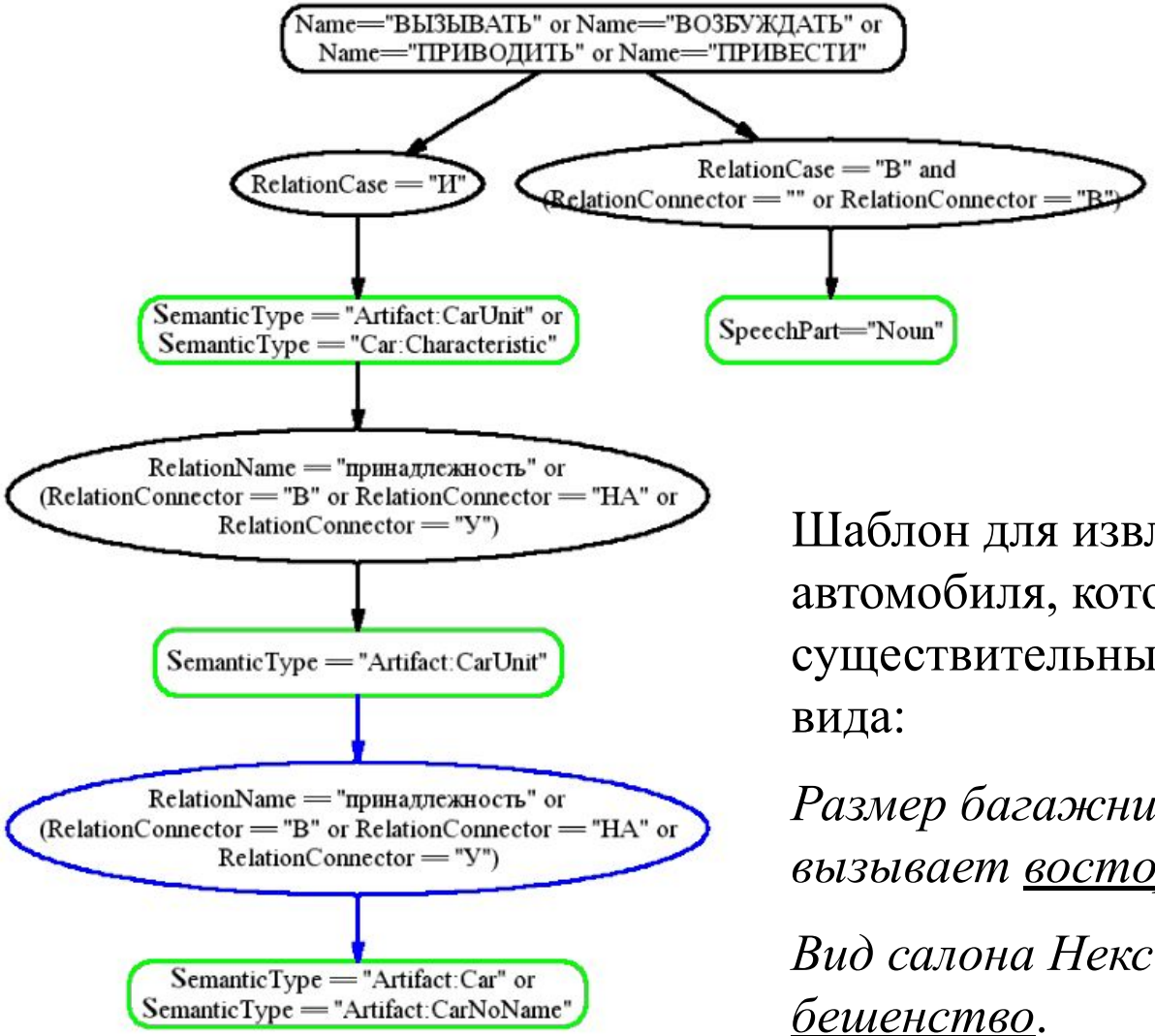



Шаблон для извлечения оценки автомобиля, которая выражается прилагательным в конструкциях вида:

*Качество у двигателя Опеля стало низким;*

*Интерьер салона в Мазде считается хорошим*.

# Извлечение знаний: семантическая интерпретация текста (2)




Шаблон для извлечения оценки автомобиля, которая выражается существительным в конструкциях вида:

*Размер багажника на Subaru вызывает восторг*;

*Вид салона Нексии приводит в бешенство*.

# Извлечение знаний из Интернета: результаты

Из 500 000 сообщений “ЖЖ” (60 Мбайт текста) извлечено:

- всего более 5000 оценок автомобилей, их узлов и характеристик;

- более 1000 (795 хороших и 328 плохих) оценок привязано к маркам автомобилей;

- более 4000 оценок узлов и характеристик не удалось привязать к конкретным маркам (связь с референтом анафорическая);

Достигнута точность: 84%

Оценка полноты: около 20%

**Спасибо за внимание!**

*Александр Ермаков, ООО «ЭР СИ О»*
*ermakov@rco.ruermakov@rco.ru, www.rco.ru*