# СИСТЕМА ПОНИМАНИЯ ТЕКСТА ДЛЯ РОБОТА Ф-2: СИНТАКСИЧЕСКИЙ АНАЛИЗ И ИЗВЛЕЧЕНИЕ СМЫСЛА[1]

**Зайдельман Л. Я.**
*zaydelman_ly@nrcki.ru*
**Котов А. А.**
*kotov_aa@nrcki.ru*
**Зинина А. А.**
*zinina_aa@nrcki.ru*
**Аринкин Н. А.**
*arinkin_na@nrcki.ru*
Курчатовский институт (Россия, Москва)

Робот Ф-2 разрабатывается для экспериментов по человеко-машинному взаимодействию, а также как прототип будущих роботов-компаньонов, способных свободно общаться с человеком. В составе робота важную роль играет компонент понимания текста: по письменному тексту на русском языке этот компонент должен восстановить семантическую структуру (смысл), чтобы на её основе робот смог выбрать ответную коммуникативную реакцию. Конструирование семантики позволяет роботу накапливать факты (пополнять знания на основе чтения текстов), а накопление фактов дает возможность уточнять структуру отдельных понятий и семантических категорий. Таким образом, в результате разбора множества предложений робот формирует аналоги обобщений над понятиями и фактами.Система понимания робота может получать текст (а) с микрофона при взаимодействии с человеком – при этом используется сторонний сервис распознавания речи, (б) из текстовых файлов – например, художественной литературы, (в) из актуальных новостей и блогов – с помощью системы RSS. Далее каждое предложение проходит несколько этаповлингвистического анализа. На морфологическом этапе каждойсловоформе приписывается лексема и список граммем: часть речи, падеж, время и т. д. Омонимичным словоформам приписываются несколько лексем или несколько наборов граммем,так, словоформа стекло может относится к существительному стеклов именительном или винительном падеже, а такжек глаголустечь. При разборе используется словарь из 100 тысяч лексем на основе проекта OpenCorpora [Грановский и др. 2010], для неизвестных слов используется система генерации гипотез (guesser). Словам также приписываются семантические признаки; в словаре они

[1] Работа поддержана грантом РФФИ 16-29-09601 «Система автоматического выявления эмоциональных и экстремистских суждений в текстах на естественном языке».

сопоставлены 30 тысячам слов (от 1 до 18 признаков на слово), например, словуматьсопоставлены признакичеловек, объект, женщина, родственник. Далее на этапесинтаксического анализа для предложения строится одно или несколько синтаксических деревьев.Используется формальное представление грамматики русского языка из 550 синтаксических правил на специально разработанном языке syntXML. Во время построения синтаксического дерева слова предложения распределяются на семантические валентности: предикат, агенс, пациенс, инструмент[Fillmore, 1968]. Семантическим представлением предложения мы считаем набор семантических признаков, распределенных по семантическим валентностям (см.таблицу 1).

Таблица 1. **Семантическое представление предложения Лингвисты отметили психологов на конференции**

| Валентность | предикат | агенс | содержание восприятия/ мышления | место |
|---|---|---|---|---|
| Признаки | думать; обращать-внимание | человек; профессия | человек; профессия | контейнер; акция; абстрактный-контейнер; абстрактный-объект |

При омонимии у предложения может быть несколько семантических представлений. Чтобы выбрать наиболее правдоподобное представление, каждый построенный смысл сравнивается с 82 семантическими сценариями [Котов, в печати]. Сценарии также состоят из признаков, распределённых по валентностям, ипприписывают тексту эмоциональную оценку: среди положительных сценариев есть сценарииприятного вкуса, комфорта, заботы, а среди отрицательных – страха, неадекватности. Семантические представления высказываний попарно сравниваются со сценариями. Наиболее близкая пара сценарий-представление «выигрывает»: представление признается наиболее подходящим данному предложению, а робот может в качестве ответа выбрать коммуникативную реакцию, связанную с выигравшим сценарием. Выбранное семантическое представление сохраняется в базе данных. Такая форма хранения данных позволяет осуществлять разноплановый поиск по семантике текстов, например, шаблону ‘некто причиняет вред кому-то, названному по национальности’ соответствуют предложения (а) как сахар убивает россиян (б) а вы могли бы убить китайца? и пр. Объединение всех контекстов с определённым словом позволяет оценивать его семантические возможности, например,

способность занимать валентности в различных ситуациях (см. таблицу 2); а объединение всех лексем, занимающих определённую валентность в ситуации, позволяет формировать семантическую категорию (см. таблицу 3).

Таблица 2. **Предложения с лексемойснег. Приведен фрагмент выдачи, иллюстрирующий богатые семантические возможности лексемы снег**

| **Валентность** | **Предложение** |
|---|---|
| инструмент | дерево полозьев облепило снегом. |
| агенс | сухой снег припорошил коричневые борозды. |
| пациенс | - *там* снег сдуло под гребнем! |
| траектория | на подъемах ямщик соскакивал с саней и шел по снегу рядом. |
| целевая точка | фонарь попал в снег, *стекло лопнуло, и свет погас.* |
| содержание восприятия или мышления | тащу и вижу: черный снег такой. |
| место | за пригорком из отвесной скалы среди снега бил ключ. |
| интерпретация | а тут еще и мелкий дождь превратился в первый снег. |
| причина | кругом все белело от снега. |
| время | после снега во сне он твердо знал, что наутро все сладится. |
| параметр или область | нам повезло со снегом. |

Таблица 3. **Объединение лексем, занимающих в предикатах типа ‘пить’ валентность пациенса. Такой запрос позволяет ответить на вопрос: ‘что обычно пьют?’В скобках указано число употреблений в базе (показаны наиболее частотные пациенсы)**

| Пациенс | Пример употребления |
|---|---|
| чай (226) | вечером все пили чай из большущего блестящего самовара. |
| вода (126) | начальники запрещали им пить воду сразу и помногу. |
| пиво (93) | как это быть грузчиком и не пить пива*!* |
| что (90) | что будем пить? |

| | |
|---|---|
| вино (88) | в ресторане сидели долго, пили вино*, шутили, смеялись, разговаривали без конца.* |
| водка (65) | водку пить, пошлости говорить? |
| кофе (57) | утром, в столовой, они пьют водянистый кофе*, жуют черствые хлебцы.* |
| кровь (47) | ты не будешь нуждаться ни в пище, ни в воде, но тебе придется пить кровь. |
| алкоголь (24) | богатые люди пьют дорогой алкоголь и меньше рискуют нарваться на подделку. |

Выделение из текста даже поверхностного семантического представления (как в таблице 1) может служить для выбора эмоциональной реакции роботом, для формирования семантического портрета отдельной лексемы (таблица 2), а также для формирования целой семантической категории (таблица 3).

*Грановский Д. В., Бочаров В. В., Бичинева С. В. 2010. Открытый корпус: принципы работы и перспективы, Компьютерная лингвистика и развитие семантического поиска в Интернете: Труды научного семинара XIII Всероссийской единой конференции «Интернет и современное общество», Санкт-Петербург,94.*

*Fillmore, C. J. 1968. The Case for Case. In Universals in linguistic theory. NewYork: Holt, Rinehart&Winston, 1-68.*

*Котов А. А. (в печати) Механизмы речевого воздействия. М.: Курчатовский институт.*