

Т. Ю. Шерстинова
T. Yu. Sherstinova

КОММУНИКАТИВНЫЕ МАКРОЭПИЗОДЫ В КОРПУСЕ ПОВСЕДНЕВНОЙ РУССКОЙ РЕЧИ «ОДИН РЕЧЕВОЙ ДЕНЬ»: ПРИНЦИПЫ АННОТИРОВАНИЯ И РЕЗУЛЬТАТЫ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ¹

COMMUNICATIVE MACRO EPISODES IN THE ORD CORPUS OF RUSSIAN EVERYDAY COMMUNICATION: ANNOTATION PRINCIPLES AND STATISTICS¹

Аннотация. В докладе представлены основные принципы аннотирования коммуникативных макроэпизодов, используемые в речевом корпусе «Один речевой день» и учитывающие тип и базовые условия повседневной речевой коммуникации; приведены некоторые результаты статистической обработки корпуса (распределение макроэпизодов по типу, месту коммуникации и социальным ролям) и выявлены наиболее типичные по данным параметрам макроэпизоды корпуса ОРД.

Abstract. The paper describes common principles for annotating communicative macro episodes in the ORD corpus of Russian everyday speech, which takes into account different types and conditions of spoken communication. The paper provides concise statistical description of the ORD macro episodes (e.g., type/place of communication and speakers' social roles), and reveals the most common types of episodes presented in the ORD corpus.

Введение

На филологическом факультете Санкт-Петербургского государственного университета при поддержке РГНФ начаты работы по созданию информационной системы коммуникативных сценариев спонтанной русской речи. Проект направлен

¹ Исследование выполнено в рамках проекта «Информационная система коммуникативных сценариев спонтанной русской речи» (грант РГНФ № 12-04-12017).

на решение фундаментальной научной проблемы описания структуры повседневной речевой коммуникации.

Речевая деятельность человека, осуществляемая им ежедневно, распадается на крупные эпизоды, объединенные местом коммуникации, ее условиями и участниками (например, «завтрак в кругу семьи», «дорога на работу», «рабочее совещание», «работа с клиентами в офисе», «обеденный перерыв», «телефонные переговоры с партнерами», «покупки в магазине», «встреча в кафе», «посещение фитнес-клуба», «домашние занятия с детьми» и т. п.). За сутки каждый человек участвует в среднем в 20-50 коммуникативных эпизодах такого типа, которые мы называем «макроэпизодами».

Речевым материалом для изучения коммуникативных макроэпизодов на русском языке является Звуковой корпус повседневной русской коммуникации «Один речевой день»². Разметка речевого материала осуществляется в программе ELAN³. Для аннотирования макроэпизодов разработана система кодирования, учитывающая 1) тип коммуникации (например, «бытовой разговор», «профессиональная беседа», «публичная речь», «клиент-сервис»-общение, «обучение/инструктаж»), 2) условия коммуникации (в частности, «кухонные» разговоры, застолья, телефонные разговоры и т. п.), 3) социальная роль говорящих (родственники: «мать», «отец», «сын», «дочь» и др., коллеги: «начальник», «подчиненный», «коллега»; друзья; клиен-

² Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation // Text, Speech and Dialogue. Proc. of the 12th International Conference, TSD 2009. Pilsen, Czech Republic. Springer: LNAI 5729, 2009. P. 250–257; Sherstinova T. The structure of the ORD speech corpus of Russian everyday communication // Text, Speech and Dialogue. Proc. of the 12th International Conference, TSD 2009. Pilsen, Czech Republic. Springer: LNAI 5729, 2009. P. 258–265.

³ ELAN – Linguistic Annotator. Version 4.5. <http://www.mpi.nl/corpus/manuals/manual-elan.pdf>

ты; знакомые; посторонние и др. категории), 4) место коммуникации (дом, офис, учебное заведение, магазин, кафе и др.).

1. Аннотирование коммуникативных эпизодов в ОРД

Аннотирование макроэпизодов осуществляется непосредственно в базе данных информационной системы и содержит следующую информацию в закодированном виде:

1.1. Тип коммуникации / коммуникативный сценарий

1) ОСНОВНЫЕ – обязательные маркеры:

П – публичная речь (лекции, презентации и т. п.).

Д – деловой разговор или профессиональная беседа, осуществляемая между коллегами или если хотя бы один из участников беседы – профессионал в обсуждаемой области.

В – общение с внешними организациями, разговор «клиент-сервис» (в магазинах, поликлиниках, библиотеках и т. п.).

У – обучение, инструктирование, практическое занятие, занятия с ребенком дома и пр.

Б – бытовой разговор в широком смысле (фактически, все типы коммуникации, не попадающие в первые 4 категории).

2) ДОПОЛНИТЕЛЬНЫЕ (факультативные) пометы. Это, в частности: К – «кухонные», З – застолья, Д – за решением профессиональных и рабочих дел, Б – за решением повседневных бытовых дел, Ж – на фоне восприятия какого-то внешнего источника, Х – «на ходу» (в дороге, на прогулке, в транспорте), Г – решение организационных, бухгалтерских, паспортных, юридических и пр. дел, Ц – осмотр врача, П – покупки и выбор товара, О – в ожидании чего-то или кого-то, Ы – отдых, «перекур», игра с ребенком, Ь – выпивая. Кроме того, специально маркируются: М – монолог (присутствие монологических фрагментов), Р – отдельные реплики и микродиалоги, Т – телефонный разговор, Ф – Фоновые разговоры, Е – Пение, Ч – чтение текстов, стихов, рассказ анекдотов и другие «цитаты», Э – позитивные эмоции, Н – негативные эмоции, конфликтные ситуации, спор, Ю – юмор, забавные ситуации.

1.2. Социальная роль информанта

Родственники: МУ – муж, ЖЕ – жена, ОТ – отец, МА – мать, СЫ – сын, ДО – дочь, СЕ – сестра, БР – брат, БА – бабушка, ДЕ – дедушка, ВН – внук, ВА – внучка.

Прочие родственники (без указания пола): СР – старшие родственники (тетя, дядя, теща, тесть, свекор, свекровь, крестный), МР – младшие родственники (зять, невестка, племянник, племянница и пр.), РР – родственники-ровесники, НР – непонятный родственник.

Друзья: БД – близкий друг (бой-френд), БП – близкая подруга (гёрл-френд), ДР – друг, ПО – подруга, АД – ассоциируемый друг (друг друзей, родственников).

Коллеги: КО – коллега, КП – коллега подчиненный (младший по возрасту), КН – коллега-начальник/командир, КД – коллега-друг или коллега-подруга.

«Парные» отношения: КЛ – клиент (в том числе – покупатель, арендатор квартиры, пациент, родители пациента), СС – сервис, сервисная служба – все, кто оказывают услуги по долгу службы или за деньги (продавец, фотограф, квартирная хозяйка, врач), УЧ – ученик, студент, ПР – преподаватель, учитель, научный руководитель, няня, ИН – интервьюирующий (журналист, следователь, ученый, спец. по кадрам и др.).

Группы людей: СТ – сокурсник, одноклассник и т. п., ПА – «партнер», ПП – «попутчик», а также посетитель того же кафе, магазина, поликлиники и пр. общественных мест, АУ – часть толпы, массовки, аудитории (на стадионе, презентации, в кино).

Прочее: ЗН – просто знакомый, ЧУ – чужой, посторонний, незнакомый человек, ХО – хозяин/хозяйка (общение с животными), СА – сам(а) себе коммуникант, включая разговоры с неодушевленными предметами.

1.3. Место коммуникации

ДОМ – дом, место проживания (квартира, общежитие, дача, номер в гостинице), ДОГ – «в гостях» (квартира, общежитие, дача, номер в гостинице), КАЗ – казарма, ОФ – офисное

(служебное) помещение, КАФ – кафе, столовая, ресторан, буфет на работе, МАГ – магазин (торговый центр, аптека, киоск), СЦ – сервис-центр (любые услуги – коммерческие и государственные учреждения, банк, ателье, библиотека, салон, офисы обслуживающих фирм, УЗ – учебное заведение (школа, университет, училище), ПОЛ – поликлиника, больница, мед. кабинет, ОМ – прочие общественные места (музей, театр, кинотеатр, храм, тир, спортзал), УЛ – улица, парк, открытый воздух, ТР – все виды транспорта, ПОД – подсобные помещения (сарай, гараж, мастерская), НО – неопределенное место.

2. Результаты статистической обработки

Объем исследованного материала – 1110 макроэпизодов⁴, полученных от 44 информантов. Общая продолжительность звучания: 288 часов.

2.1. Распределение по типу коммуникации

Более половины всех макроэпизодов корпуса (57%)⁵ отражают бытовую коммуникацию. Деловые (профессиональные) разговоры составляют 21% всех эпизодов. Коммуникация по типу «клиент-сервис» насчитывает 10%, 6,5% относятся к учебному процессу, а 3% составляют публичные выступления.

В свою очередь, наиболее частотные бытовые разговоры имеют место за решением бытовых дел (11%), представляют собой отдельные реплики (9%) или разговоры за несправочной едой (9%), и осуществляются на ходу в 7,5% случаев.

Деловые или производственные разговоры происходят как правило за решением производственных дел (10%), 3% всех макроэпизодов такого типа активно используют монологическую речь и такой же процент составляют отдельные реплики.

⁴ Аннотирование макроэпизодов выполнено С. Б. Степановой и Т. Ю. Шерстиновой.

⁵ Здесь и далее проценты приводятся к общему количеству всех макроэпизодов.

Коммуникация по типу «клиент-сервис» на материале ОРД лучше всего представлена следующими видами макроэпизодов: ВЦ – осмотр или консультация врача (3,8%), ВП – выбор товара и покупки (2,3%), ВУ – инструктаж или консультирование (0,7%).

2.2. Распределение социальных ролей

В таблице ниже представлен ранжированный список наиболее частотных социальных ролей, в которых выступают информанты.

Социальная роль	%	Социальная роль	%
коллега	22%	знакомый	2%
подруга	7,2%	коллега-друг	2%
преподаватель	6%	студент	2%
сервис-служба	6%	сам(а) с собой	1%
сокурсник	4%	отец	1%
муж	3%	попутчик	1%
клиент	3%	мать	1%
друг	3%	партнер	1%

Наиболее типична (1/5 часть всех разговоров корпуса) роль «коллеги». На втором месте по активности выступают «подруги». Преподаватели (учителя) и работники сервисных служб поделили третье место. Не представленные в таблице социальные роли информантов встретились на материале ОРД менее чем в 1% случаев, тем не менее суммарно они составляют 34,5%.

2.3. Распределение мест коммуникации

В следующей таблице представлено распределение макроэпизодов корпуса ОРД по месту коммуникации. Как и следовало ожидать, максимальное количество разговоров осуществляется «дома» (точнее, в месте постоянного проживания) и на работе (в офисе). Кроме того, в корпусе хорошо представлена коммуникация в учебных заведениях, «в гостях» и сервисных службах, а также в медицинских центрах и на улице.

Место коммуникации	%	Место коммуникации	%
дом	18%	др. обществ. место	5%

офис	16%	кафе	5%
учебное заведение	12%	транспорт	5%
в гостях	8%	магазин	4%
сервис-центр	7%	казарма	2%
на улице	6%	подсобн. помещение	1%
поликлиника	6%	прочее	5%

2.4. Наиболее типичные макроэпизоды в зависимости от социальной роли информанта и места коммуникации

Автоматическая обработка аннотаций позволила получить все встречающиеся в корпусе комбинаторные варианты сочетаний социальной роли информанта и места коммуникации. Их оказалось 344, для каждой комбинации были получены частотные таблицы.

Наиболее типичными комбинациями корпуса ОРД с этой точки зрения оказались следующие:

1. **Коллега – Офис** (125 эпизодов, 11%)
2. **Преподаватель – Уч. Заведение** (47 эпизодов, 4%)
3. **Подруга – Дом** (35 эпизодов, 3%)
4. **Коллега – Сервис-Центр** (34 эпизода, 3%)
5. **Муж – Дом** (26 эпизодов, 2%)

Абсолютная частота некоторых других факультативных помет: М (монологические фрагменты) – 298, Т (телефонный разговор) – 190, К («кухонные» разговоры) – 146, Ю (юмор) – 138, Б (выпивают) – 38, Э (положительные эмоции) – 48, Н (негативные эмоции) – 32, З (застолья) – 13.

Кроме того, для каждого из информантов был проанализирован спектр разных условий коммуникации. Наиболее разнообразный материал показали следующие информанты:

- S24 – 12 разных ролей / 7 типов мест
- S35 – 10 разных ролей / 6 типов мест
- S30 – 10 разных ролей / 8 типов мест
- S04 – 9 разных ролей / 7 типов мест
- S11 – 48 типов эпизодов / Б Д У В
- S05 – 23 типа эпизодов / Б Д У В

Заключение

Полученные статистические характеристики по корпусу ОРД, разумеется, далеко не во всем отражают коммуникативную ситуацию в обществе: в первую очередь это характеристики речевого материала на текущий момент создания корпуса, которые могут быть использованы, в частности, для выявления путей дальнейшей балансировки речевого материала. Описанное аннотирование речевого материала и результаты его обработки будут использованы для разработки поисковой системы по данным параметрам для коммуникативных исследований, а также для анализа речи, характерной для того или иного типа и условий коммуникации.