

Е. А. Гришина

Мультимедийный русский корпус (МУРКО):

проблемы аннотации¹

1.

```
1. Данный текст был распознаны в с...
2. ...на даёмых... "мечта в это же...
3. "с. всё-таки" ...привыкла >
4. ...но... что дивится
5. ...и... почему же
6. ...и так был ...достоин же
7. ...ны... а несомненно
8. ...на... а невольно
9. ...на... а невольно
10. ...на... а невольно
11. ...на... а невольно
12. ...на... а невольно
13. ...на... а невольно
14. ...на... а невольно
15. ...на... а невольно
16. ...на... а невольно
17. ...на... а невольно
18. ...на... а невольно
19. ...на... а невольно
20. ...на... а невольно
21. ...на... а невольно
22. ...на... а невольно
23. ...на... а невольно
24. ...на... а невольно
25. ...на... а невольно
```

работе [Гришина, Савчук 2008] было дано общее описание проекта Корпуса звучащей русской речи, который к настоящему моменту получил название «Мультимедийный русский корпус» (мурко)². Поэтому в данной статье мы лишь кратко охарактеризуем этот проект и основное внимание уделим проблемам возможной разметки корпуса мурко (очевидно, что разработка принципов разметки корпуса практически равна разработке параметров поискового интерфейса, к которому, как предполагается, будет обращаться пользователь).

Планируется, что мурко будет создаваться прежде всего на кинематографическом материале, накопленном к настоящему моменту

1 Исследование проведено при поддержке грантов РФФИ 06-06-80133-а и 08-06-00371-а, а также программы ОИФН РАН «Генезис и взаимодействие социальных, культурных и языковых общностей».

2 Предварительный анализ возможности создания МУРКО был проведен при очень важной содержательной поддержке программистов компании «Яндекс» Андрея Аброскина и Николая Григорьева, за что автор выражает им глубокую благодарность.

в ходе пополнения подкорпуса устной речи и акцентологического подкорпуса в рамках нкря. По мере развития мурко в него будут включены также те записи устной речи из устного подкорпуса, для которых имеются в наличии соответствующие видео- и аудиофайлы.

Принцип построения мурко достаточно очевиден—файлы фильмов, видео- и аудиофайлы будут разрезаны на минимальные целостные единицы (клипы), и им в соответствие будут поставлены элементы транскриптов. Те клипы, которые содержат некоторый жестовый материал, но не содержат никакого текста, будут включены в мурко, естественно, без соответствующих скриптов. Клипы, сопровождающиеся текстовым материалом, в дальнейшем будут называться *кликстами* (или *клипотекстами*), а видеофрагменты, содержащие только жестовый материал,—собственно *клипами*.

Таким образом, единицей выдачи в мурко будут 1) для кликстов—фрагменты транскриптов, размеченные морфологически и семантически по методике, принятой в нкря, и связанные гиперссылкой с соответствующим клипом; 2) для клипов—гиперссылки, отсылающие к базе данных, содержащей набор клипов.

Каждый кликст/клип будет расцениваться как отдельный текст, имеющий свое метаописание (так, как это принято в нкря). Это метаописание будет создано в соответствии с набором параметров, разработанных для устного подкорпуса нкря (параметры, связанные с автором текста как целого, датой создания, жанровой характеристикой и т.д., а также параметры, относящиеся к социологической разметке,—автор реплики (для фильмов—имя персонажа и актера), год его рождения и пол, если таковые известны). Следовательно, как и в нкря, в соответствии с этими параметрами (с каждым в отдельности или с их комбинацией) пользователь сможет формировать свой собственный подкорпус.

Уже в таком виде мурко будет представлять интерес для пользователя, особенно в условиях дефицита мультимедийных корпусов, созданных на материале русского языка. Появится возможность получить звуковой и/или жестовый материал, ведя его поиск 1) от леммы/словоформы, 2) от словосочетания, 3) от морфологических характеристик, 4) от семантических характеристик, 5) от имени персонажа, 6) от имени актера, 7) от возраста и пола говорящего,

8) от сочетания всего перечисленного. Поскольку предполагаемый объем только кинематографического материала—порядка 3 млн словоупотреблений, даже при такой минимальной разметке мурко представляет большой интерес для исследователей.

Так, например, уже на этом материале мы можем ставить и решать некоторые проблемы, связанные с фонетикой/орфоэпией. Например, сформулировав запрос от лексем с корнем *бухгалтер* (*бухгалтер** в основной строке поиска), мы получаем следующую выдачу³.

Таблица 1

Фильм	Дата создания фильма	Актер	Дата рождения актера	Способ произнесения
Беспокойное хозяйство	1946	Александр Граве	1920	Фрикатив <i>буауалтер</i>
Мы с вами где-то встречались...	1954	Владимир Лепко	1898	
Мы с вами где-то встречались...	1954	Андрей Тутышкин	1910	
Укротительница тигров	1954 (2 раза)	Павел Суханов	1911	
Укротительница тигров	1954	Людмила Касаткина	1925	
Укротительница тигров	1954 (2 раза)	Константин Сорокин	1908	
Дети Дон Кихота	1965	Наталья Фатеева	1934	
Ко мне, Мухтар!	1965			
Вас вызывает Таймыр	1970	Евгений Весник	1923	
Солярис	1972	Владислав Дворжецкий	1939	
Люди и манекены	1974 (3 раза)	Аркадий Райкин	1911	
Люди и манекены	1974			
Театр	1978			

³ Анализ проводился на корпусе объемом 1,2 млн словоупотреблений (так, как он сложился на июнь 2008 г.).

Фильм	Дата создания фильма	Актер	Дата рождения актера	Способ произнесения
Дети Дон Кихота	1965	Николай Парфенов	1912	Смычный бугалтер
Берегись автомобиля	1966			
Белорусский вокзал	1970	Николай Волков	1934	
Вас вызывает Таймыр	1970	Юрий Кузьменков	1941	
Театр	1978 (4 раза)			
Родня	1981	Геннадий Бортник	1939	

Анализ материала показывает, что в среднем фильм «с фрикативом» на 13 лет старше фильма «со смычным», а актеры, произносящие в данном корне фрикатив, в среднем на 14 лет старше актеров, произносящих смычный:

Таблица 2

	Средний год создания фильма	Средний год рождения актера
Фрикатив	1960	1917
Смычный	1973	1931

Следовательно, произнесение с фрикативом может расцениваться как старшая, а со смычным—как младшая норма.

Еще один пример. Сформулировав запрос «сочетание предлога к со словом, начинающимся с к-», мы получим материал, фрагмент которого представлен в Таблице 3⁴.

⁴ Приведена лишь часть материала, доступного в кинематографическом под-корпусе на июнь 2008 г., поскольку в отсутствие мультимедийного корпуса поиск соответствующего клипа представляет довольно большие сложности. Впрочем, для иллюстрации возможностей мурко ни объем материала, ни даже правильность сделанных на основании отобранного материала выводов не имеет большого значения—важно подчеркнуть перспективы использования мурко.

Таблица 3⁵

Фильм	Дата	Сочетание	Актер	Пол актера	Дата рождения	Способ произнесения
Встреча на Эльбе	1949	к капитану	Борис Андреев	м	1915	гемината
Встреча на Эльбе	1949	к коммунизму		м		гемината
Мы с вами где-то встречались	1954	к кому	Аркадий Райкин	м	1911	гемината
Мы с вами где-то встречались	1954	к классике	Аркадий Райкин	м	1911	одиночное К
Добровольцы	1958	к кому	Элина Быстрицкая	ж	1928	одиночное К
Берегись автомобиля	1966	к концу	Юрий Яковлев	м	1928	одиночное К
Вас вызывает Таймыр	1970	к Кирпичникову	Юрий Кузьменков	м	1941	одиночное К
Старик-разбойник	1971	к кошмару	Юрий Яковлев	м	1928	гемината
Мачеха	1973	к камере		м		гемината
Мачеха	1973	к каким	Татьяна Доронина	ж	1933	одиночное К
Люди и манекены	1974	к кому	Аркадий Райкин	м	1911	гемината
Приключения Буратино	1975	к Карабасу		ж		одиночное К
Приключения Буратино	1975	к Карабасу		ж		одиночное К
Театр	1978	к концу		м		гемината
Адам женится на Еве	1980	к каким	Татьяна Васильева	ж	1947	одиночное К
Адам женится на Еве	1980	к которм	Александр Калягин	м	1942	одиночное К
Будьте моим мужем	1981	к кому	Елена Проклова	ж	1953	одиночное К

⁵ Зафиксировано также три случая диссимилиции (персонаж Зиновия Герда в фильме «Адам женится на Еве», Евгения Леонова в фильме «Дом, который построил Свифт» и Сергея Гусинского в фильме «Операция «С Новым годом!»»).

Фильм	Дата	Сочетание	Актер	Пол актера	Дата рож- дения	Слобод произнесе- ния
Родня	1981	к кому	Андрей Петров	м	1919	одинокое К
Дом, который построил Свифт	1983	к камину	Владимир Белоусов	м	1947	гемината
Дом, который построил Свифт	1983	к камину	Александра Заха- рова	ж	1962	одинокое К
Двадцатый век начинается	1986	к концу		м		одинокое К
Московские каникулы	1995	к командиру	Ирина Селезнева	ж	1961	одинокое К
Операция «С Новым годом!»	1996	к кому	Ирина Полянская	ж	1967	гемината
Операция «С Новым годом!»	1996	к кому	Андрей Краско	м	1957	гемината
Операция «С Новым годом!»	1996	к Качалову		ж		одинокое К
Ландыш серебристый	2000	к кому	Алена Хмельницкая	ж	1971	одинокое К
Ландыш серебристый	2000	к краю	Валерий Гаркалин	м	1954	одинокое К

Анализ материала по параметру возраста актера и года создания фильма дает менее отчетливую картину, чем в предыдущем примере:

Таблица 4

	Средний год созда- ния филь- ма	Средний год рож- дения ак- тера
Гемината	1963	1934
Одиночное К	1984	1941

Как видим, разница в возрасте актеров (1934–1941) слишком мала, чтобы относиться к ней серьезно. Разница в годе создания фильма—довольно существенна (21 год). Таким образом, мы здесь сталкиваемся либо с недостатком материала, либо со своеобразным соотношением старшей и младшей нормы—когда они связаны скорее с эпохой создания фильма, чем с речевой манерой актера.

Более отчетливое распределение дает анализ по полу актера:

Таблица 5⁶

	Все	Мужчины	Женщины и дети
Гемината	37%	56%	9%
Одиночное К	63%	44%	91%

Как видим, введение параметра пола дает значительные отклонения от среднего распределения, а следовательно, пол говорящего в данном случае должен рассматриваться как существенный параметр (если говорить в общем, мужчины предпочитают напряженное произнесение, а женщины и дети—расслабленное).

Подтвердятся ли эти закономерности при расширении материала, и если подтвердятся, то каковы соотношения между хронологическим и гендерным аспектом в употреблении данных вариантов,—это уже вопрос не к корпусу, а к тем, кто будет им пользоваться профессионально. Важно подчеркнуть, что мурко дает возможность ставить такие вопросы.

⁶ В таблице полужирным даются данные, существенно превосходящие средние значения, а курсивом, напротив, данные, существенно уступающие средним значениям.

Очевидно, что даже с такой минимальной разметкой мурко будет исключительно важен и для исследователей русской интонации—в тех ее проявлениях, которые связаны с лексикой, морфологией, семантикой и социологией (гендерными и возрастными характеристиками говорящих), а также для исследователей русской системы жестов (в аспекте связи жеста со словом).

2.

Представляется, однако, что имеющаяся аннотация, полностью заимствованная из НКРЯ, в случае мурко должна рассматриваться как необходимая, но далеко не достаточная. Требуется разработать некоторую систему разметки, которая позволяла бы обращаться к материалу мурко вне зависимости от слова во всех его ипостасях (как такового, как имеющего морфологическую, семантическую и социологическую характеристику).

Это, в сущности, обозначает, что должно быть существенно, принципиально расширено метаописание кликста/клипа, с тем чтобы у пользователя появилась возможность отбирать однородные в том или ином отношении кликсты/клипы вне зависимости от их словесного наполнения⁷. Кроме того, ясно, что только расширение метаописания клипов позволит обращаться к жестовому материалу в том случае, если жестовое содержание клипа не сопровождается лексическим рядом (в кинематографе эти случаи, очевидно, достаточно частотны), или в том случае, если словесное содержание кликста является, в сущности, несловесным (содержит то, что в английской традиции называется *nonverbal words*)—включает, например, междометия или вокальные жесты (см. [Шаронов 2008]). Кроме того, как известно, ряд самых разных по прагматическим характеристикам междометий традиционно кодируются на письме одним и тем же способом (например, *А!* разочарования, понимания–узнавания и нек. др.), и единственный способ различить их—приписать кликсту в целом ту или иную ситуацию употребления данного междометия или вокального жеста, которая для

⁷ Как станет ясно из дальнейшего изложения, такое расширение метаописания в конечном итоге может повысить поисковые возможности и самого НКРЯ.

разметчика кликста достаточно очевидна, а для обычного пользователя, имеющего в своем распоряжении только транскрипт, иногда непонятна даже и при максимальном расширении контекста.

3.

Расширение стандарта метаописания для устных или мультимедийных (в англо-американской традиции—мультимодальных, *multimodal*) корпусов—общее место современной корпусной лингвистики и инженерии.

Поскольку перед европейской и американской корпусной лингвистикой стоят вполне конкретные задачи максимального приближения общения человека с компьютером к стандартам естественного общения человека с человеком, то перед ней в полный рост встает проблема определения тех параметров, которые, собственно, и могут способствовать этому приближению⁸.

Эти параметры группируются следующим образом: 1) построение типологии речевых действий (*dialogue acts=speech acts=dialogue moves*), которые стандартно употребляются в тех или иных ситуациях⁹ говорящим; создание корпусов, ставящих перед собой такую задачу, позволяет в конечном итоге определить состав слов и словосочетаний, а также интонационных контуров, характерных для тех или иных речевых действий, с тем чтобы машина в будущем могла однозначно и правильно реагировать на такого рода сигналы; 2) построение типологии человеческих эмоций

⁸ Как показала последняя международная конференция по Language Resources and Evaluation LREC'2008 (Марокко, Марракеш, 28-30 мая 2008 г.), создание и аннотация мультимедийных корпусов—мейнстрим современной мировой корпусной лингвистики, имеющий, однако, в качестве своей доминанты не теоретическое изучение тех или иных языков в мультимедийном аспекте, а вполне конкретные инженерные задачи (см. материалы конференции на специальном сайте [LREC'2008], к которому мы и будем отсылать в дальнейшем).

⁹ Ситуации могут быть самыми разнообразными – заказ блюд в ресторане ([Strauß et al. 2008]), общение с системой «умный дом» ([Möller et al. 2008]; [Georgila et al. 2008]; [Kostoulas et al. 2008]), с телевизором в интерактивном режиме ([Brutti et al. 2008]), общение с городским транспортным центром ([Marasek, Gubrynowicz 2008]), корпус обращений в американскую службу 911 на языках, отличных от английского ([Nallasamy et al. 2008]), и т.д.

и настроений, чтобы машина могла учитывать в своей реакции не только содержательную, но и эмоциональную составляющую человеческого запроса¹⁰; 3) корпусное исследование тематического развития диалога, включая исследование проблем референции¹¹; 4) построение корпусов жестов—как самоцель, либо как часть мультимедийных корпусов (авторы—вполне логично—исходят из того, что полноценное общение человека с машиной предполагает правильную реакцию машины не только на вербальную, но и на жестовую информацию)¹².

Одновременно следует отметить два типа экстралингвистических ограничений, которые накладываются в Европе и США на создание мультимедийных корпусов. Во-первых, это ограничения в финансировании—мультимедийные корпуса создаются большей частью в рамках тех или иных коммерческих или социальных проектов¹³ (что естественно связано с конкретной постановкой задачи в ходе построения этих корпусов), соответственно, финансируются вполне конкретные и исчислимые заранее результаты. Это очевидным образом ведет как к разумному ограничению объемов корпусов, ни один из которых не достигает и одного миллиона словоупотреблений (обычно же мультимедийный корпус в 100 000 словоупотреблений считается очень большим), так и к ограничениям на тип общения, на основе которого сделан тот или иной корпус. Во-вторых, существуют жесткие и, по-видимому, пока непреодолимые ограничения, связанные с нарушением *privacy* «испытуемых» и с правами на воспроизведение полученных материалов (копирайт).

¹⁰ Выступлений, касающихся построению корпусов эмоций, на LREC'2008, было достаточно много (две секции в ходе основной конференции и два специальных семинара перед конференцией). Назовем только небольшую часть: [Forbes-Riley et al. 2008]; [Gnjatović, Rösner 2008]; [Wilson 2008]; [Devillers, Martin 2008]; [Sainz et al. 2008]; [Fék et al. 2008]; [Cullen et al. 2008]).

¹¹ См. [van Son et al. 2008]; [Stoia et al. 2008]; [Gallo et al. 2008]; [Wilks et al. 2008].

¹² Аннотация направления и длительности взглядов – [van Son et al. 2008], жестов рук – [Savino et al. 2008]; комплексная аннотация мимики и жестов – [Knight, Tennent 2008]; [Blache et al. 2008]).

¹³ Например, проект, финансируемый ЕС и предполагающий адаптацию пожилых людей к пользованию современной техникой (системой «умный дом», смартфонами, КПК, коммуникаторами и проч.).

Эти ограничения ведут либо к недоступности создаваемых корпусов для широкого пользователя («на публику» могут выноситься лишь научные и инженерные наработки, сделанные на материале этих корпусов, что в значительной степени обедняет собственно лингвистические возможности их использования), либо к искусственности получаемого материала (например, используется разыгрывание тех или иных ситуаций с помощью нанятых актеров или моделируется естественное поведение «испытуемых» в тех или иных условиях с помощью компьютерных симуляторов в режиме Wizard of Oz (WOZ), когда «испытуемый», общаясь с человеком, думает, что общается с машиной; для этих же целей используются и компьютерные игры типа Quake). В редких случаях создателям корпуса приходилось обзаводиться договорами, в которых гарантируется будущая «неподсудность» создателей корпуса и отсутствие претензий со стороны «испытуемых» (см., например, [van Son et al. 2008]).

Каковы на этом фоне характеристики будущего русского мультимедийного корпуса?

1. Поскольку мурко планируется создавать в рамках нкря, то базовые экстралингвистические характеристики нкря—его открытость для всех и академический, а не коммерческий характер его использования, «вписанный» в саму его структуру,—будут свойственны и мурко.

2. Как и нкря, мурко будет построен на принципах цитирования—то есть из выдаваемых в нем контекстов (кликстов/клипов) никоим образом нельзя будет получить целый текст (фильм или видеозапись). Тем самым будут соблюдены авторские права (напомним, что именно это свойство выдачи контекстов в нкря позволило расширить корпус до 160 млн словоупотреблений, из которых более 90 млн относятся ко второй половине 20-го и началу 21-го века,—и издатели, и авторы, и научные коллективы, предоставлявшие нкря тексты, были гарантированы от неконтролируемого распространения их книг и материалов в Интернете и контрафактного переиздания, будь то на бумаге или в электронном виде).

3. Поскольку мурко планируется создавать прежде всего на базе советского и российского кинематографа, а также на материале теле- и видеозаписей, то не возникает проблемы нарушения privacy.

4. Бесконечное разнообразие ситуаций, отраженных в кинематографе, позволяет ставить и решать лингвистические, общелингуистические, психологические, исторические и прочие задачи на самом разнообразном материале, включая ситуации, которые вряд ли в принципе доступны для создателей *real-life* корпусов (сомнительна не только возможность произвести студийную запись текста в ходе, например, боевых действий, но и в гораздо более мирных, гражданских условиях—например, когда человек внезапно падает или когда он зовет на помощь, находясь в полном одиночестве). Тем самым широкому пользователю предоставляется гораздо более разнообразный материал, чем это возможно при работе с *real-life* корпусами.

5. При благоприятных условиях мурко, как уже говорилось выше, в конечном итоге будет иметь объем не менее 3 млн словоупотреблений, что делает статистические данные, полученные на его материале, гораздо более достоверными, чем статистические данные, полученные на основе корпусов объемом 0,1 млн словоупотреблений, не говоря уже о меньших объемах.

6. Из четырех направлений аннотации мультимедийных корпусов, перечисленных выше, для мурко актуальными, по нашему мнению, являются следующие: 1) аннотация типов речевых действий и 2) аннотация жестов. Именно эти два типа аннотаций имеют под собой более или менее объективные основания для классификации (см. об этом ниже). Что касается аннотирования тематического развития текста, а также аннотирования эмоций, то здесь отсутствие формальных критериев для классификации может привести к тому, что, во-первых, разные разметчики будут приходить к разным результатам, а во-вторых, представления *пользователя* корпуса о том, какая именно тема развивается в данном кликсте и какую эмоцию выражает в данный момент тот или иной говорящий, может не совпасть с представлениями *разметчика*, что приведет к системным сбоям в получении информации из корпуса. Таким образом, базовыми параметрами при разметке мурко будет разметка речевых действий и разметка жестов¹⁴.

7. И наконец, мурко может послужить базой для создания Учебного мультимедийного русского корпуса (умко), основу которо-

¹⁴ Как будет ясно из дальнейшего, при разметке жестов возможно будет учесть ту или иную эмоциональную составляющую высказывания.

го могли бы составить мультфильмы и детские фильмы, входящие в мурко. Учебный мультимедийный корпус мог бы найти применение при обучении русскому языку как иностранному (впрочем, вполне вероятно использование умко и для обучения русскому языку как родному). В рамках специализированных корпусов эта задача не решаема—здесь возможно только создание корпусов, предназначенных для тренинга автоматических обучающих систем, см., например, об этом [Forbes-Riley et al. 2008].

Таким образом, мурко планируется создавать в качестве **национального** мультимедийного корпуса, аналоги которого в мировой практике нам пока не известны.

Основной претензией, единственным «но» при создании корпуса на основе кинематографических данных является тот факт, что устная речь здесь не может рассматриваться как спонтанная. Именно в связи с этим и в выступлениях на LREC'2008, и в частных беседах с участниками конференции при обсуждении перспектив использования кинематографической речи в качестве базы для создания мультимедийного корпуса предпочтение отдавалось корпусам т.н. спонтанной речи (при том, что реальная спонтанная речь создателям корпусов либо недоступна, и ее приходится так или иначе симулировать, либо ее объем слишком ограничен и привязан к очень узкому кругу ситуаций, либо корпуса, содержащие спонтанную речь в значительном объеме, недоступны для широкого использования и ни в коей мере не могут претендовать на статус национальных, см. об этом выше). В частности, одним из аргументов против использования кинематографа для построения мультимедийного корпуса является тот факт, что при разыгрывании тех или иных эмоций нанятыми актерами довольно низкий процент эмоций правильно опознавался сторонними пользователями. Кроме того, некоторые исследователи обращают внимание на то, что «...существует большой разрыв между эмоциональными состояниями, наблюдаемыми в искусственных условиях (разыгранные сцены или искусственно стимулированные проявления эмоций), и эмоциональными состояниями, наблюдаемыми в естественном повседневном спонтанном общении» [Devillers, Martin 2008]. В частности, этот разрыв сказывается в том, что разыгранные эмоции проявляются гораздо более аффективно, чем эмоции в реальной жизни.

Однако существует и другая точка зрения. Некоторые исследователи указывают, что дефект не в актерском розыгрыше эмоций как таковом, а в неверной методологии: актерам предлагаются для произнесения отдельные предложения, вырванные из содержательного, событийного и эмоционального контекста, и в этих условиях, естественно, актер должен весь контекст «вложить» «в единое слово», что и ведет к чрезмерной аффектации ([Busso, Narayanan 2008]). «Вместо монологов и коротких предложений база данных [для исполнения актером] должна содержать естественные диалоги, в которых эмоции проявляются естественным и уместным образом» (ib.). Более того, средняя длительность диалогов должна быть достаточной для того, чтобы послужить контекстом для физических проявлений эмоций и протекания эмоционального состояния (ib.). Очевидно, что кинематограф полностью удовлетворяет этим условиям. Более того, именно кинематограф, по сравнению, например, с театральным представлением, позволяет максимально естественно разыграть эмоции, поскольку для спектакля, как известно, необходимо подчеркнуть довольно высокую рампу между сценой и зрительным залом, а в кинематографе эта граница сама по себе достаточно отчетлива и не нуждается в специальном акцентировании.

Все, что было выше сказано об эмоциях в кинематографе, относится в целом к естественности речи в кино (нам уже приходилось писать об этом—[Гришина 2007], [Grishina 2007], [Гришина, Савчук 2008], в частности, со ссылкой на новаторскую работу [Капанадзе 1986]). Добавим только следующее—как показатель, так сказать, вторичной непринужденности речи актера в кино (т.е. непринужденности, которая является не результатом спонтанности речи, а результатом настолько полного усвоения актером чужой речи, что она становится как бы полностью его собственной) могут восприниматься речевые ошибки и оговорки, которые остаются в результирующем «тексте» фильма. Если актеру удавалось поймать необходимую интонацию, произнести текст роли абсолютно органично, то режиссеры не считали нужным переозвучивать сцену, содержащую речевые ошибки и оговорки. Например, в фильме Г. Данелии «Кин-дза-дза» нет ни одного случая употребления глагола *надевать* с винительным падежом неодушевленного суще-

ствительного—используется только глагол *одевать*, хотя в фильме играют такие актеры, как Станислав Любшин, Евгений Леонов и Юрий Яковлев, которых трудно заподозрить в недостаточном уровне речевой культуры. Оговорка Евгения Леонова в фильме Марка Захарова «Обыкновенное чудо»—*Простите мне такую тонкость... грубость выражений*—сохранена в звуковой дорожке фильма (и именно в таком виде, с оговоркой, отсутствующей в пьесе Е.Шварца, эта фраза стала крылатым выражением). В фильме «Адмирал Ушаков» Борис Ливанов, игравший Потемкина, оговаривается и произносит *Неужто Пугачев душегуба посмел на корабле укрыть?* вместо *Неужто Ушаков душегуба посмел на корабле укрыть?* В фильме «Операция „С новым годом!“» одна из героинь использует форму *взяна* вместо *взята*, а её собеседник использует местоимение *она* вместо правильного *он*. И так далее, эти примеры можно множить. Все это, как представляется, свидетельствует в пользу того, что актерскую речь в фильмах можно расценивать как чрезвычайно приближенную к естественной и спонтанной.

4.

Теперь необходимо более подробно изложить способы дополнительной метаразметки в мурко, о которой говорилось выше. Начнем с метаразметки *речевых действий*, содержащихся в кликстах.

Поля, по которым размечается речевая составляющая клиста¹⁵:

- 1) социальная ситуация
- 2) речевое действие
- 3) полнота речевого действия
- 4) манера говорения
- 5) типы повторов
- 6) количество говорящих
- 7) язык

¹⁵ Типы аннотации, которые предусматриваются для дополнительной разметки речевой и жестовой составляющей клиста/клипа, проработаны на данный момент в разной степени. Что касается основных полей разметки, то они, по-видимому, установлены окончательно. Конкретное же наполнение каждого поля, напротив, находится в стадии разработки и, более того, будет уточняться (пополняться, укрупняться и под.) и далее, в том числе и в ходе создания самого корпуса.

- 8) невербальная составляющая (междометия, вокальные жесты, физиологические действия)

Под **социальными ситуациями** имеются в виду типичные социальные ситуации, в которые попадает человек,—ситуации, в которых общение наиболее формализовано, отлито в устойчивые формы. Предварительно вычленяются: *журналистский репортаж, заказ такси, застольная речь, знакомство, интервью, конференс, заказ в ресторане, разговор с водителем такси, телефонный разговор, разговор с представителем власти, разговор с продавцом, рассказ экскурсовода, урок, выступление на собрании, выступление на митинге*. Этот ряд открыт, но, по-видимому, не бесконечен и будет пополняться по мере описания новых фильмов¹⁶. Отметим, что поле «ситуация»—общее для кликстов и для клипов, т.е. возможно и необходимо приписывание некоторой ситуации тому или иному клипу даже в том случае, если он не сопровождается словесным рядом.

Под **речевыми действиями** понимается то или иное речевое действие, которое совершается в данном кликсте его участниками. Одному и тому же кликсту может быть приписано несколько речевых действий. Этот момент требует отдельного пояснения. Прежде всего, говорящий (говорящие) на протяжении одного кликта могут произвести несколько разных речевых действий (например, *вопрос, ответ, предложение, согласие*). Однако множественность описания одного кликта по данному параметру связана не только с этим фактом. Для прояснения ситуации следует объяснить, как именно мы вычленяем те или иные речевые действия.

В имеющейся мировой практике есть, по-видимому, два принципиально разных подхода к вычленению речевых действий. Первый—логический, исходящий из оптимистической предпосылки, что существует возможность априори исчислить все типы речевых действий, которые способен произвести человек. Этот принцип используется, в частности, в классификации DAMSL, с помощью которой, среди прочего, был аннотирован один из самых крупных общедоступных корпусов устной речи SWITCHBOARD (DAMSL включает 50 классов речевых актов, которые вместе с дополнительными параметрами предлагают для разметки 220 речевых актов более

¹⁶ Предварительный анализ дополнительных полей метаразметки проведен на материале фильмов «Бриллиантовая рука» и «Друг мой, Колька».

низкой ступени иерархии). Несколько иной взгляд на проблему предполагает, что классификация речевых актов должна быть эмпирической, т.е. считается принципиально невозможным априорное исчисление всех возможных речевых действий (см., например, [Hennoste et al. 2008]).

Мы считаем, что второй подход более продуктивен. Во-первых, при кажущейся объективности первого—логического—подхода, практически невозможно добиться, чтобы разные разметчики могли одинаково применить одно и то же логическое понятие в разных эмпирических случаях, тем самым мы обречены на неуправляемый разноречивой разметке. Во-вторых, практика применения логических иерархий речевых актов для разметки реальных корпусов показывает, что максимально хорошие результаты¹⁷ достигаются в тех случаях, когда принятая система понятий наиболее естественна и интуитивно понятна (см., например, об этом [Geertzen et al. 2008])¹⁸.

В связи с вышеизложенным, нами было принято решение при разметке речевых действий обращаться в большинстве случаев к тем типам речевых актов, которые отражены **в самом русском языке в совокупности русских глаголов речи** (с минимальным необходимым привлечением чисто лингвистических понятий, например, *общий и частный вопросы*). Это дает нам шанс на то, что при аннотации два разных разметчика будут выдавать максимально близкие результаты, и, кроме того, пользователь, который владеет тем же русским языком, что и разметчики, будет интуитивно понимать, что именно имелось в виду при разметке, когда, например, употреблялась метка *требование*. Кроме того, положительный момент в таком имманентном языке способе разметки заключается в том, что множества объектов, названные близкими, но разными

¹⁷ Максимально хорошими считаются результаты, при которых достигается максимальная согласованность полученных результатов при разметке одного и того же материала разными разметчиками, например, экспертами и неподготовленными разметчиками.

¹⁸ Интересно, что строгая логическая иерархия речевых актов при работе с реальными текстами максимально упрощается («сплющивается») — т.е. из нее берутся только сущности низших ступеней иерархии, максимально приближенные к реальным текстам (так, исходные 220 логических сущностей DAMSL были «упрощены» и «сплющены» до 42-х реально работающих меток (см. об этом [Webb et al. 2008])).

глаголами речи (вернее, их именными производными), пересекаются, а сами глаголы речи выстраиваются в группы. Тем самым, например, если один разметчик обозначил некоторое событие с помощью метки *требование*, а второй использовал для этого же явления метку *распоряжение*, то у пользователя есть шанс получить искомое событие, запросив *требование* и/или *распоряжение*. Таким образом, при такой разметке присвоенные речевому событию наименования как бы подстраховывают друг друга.

Именно этот имманентный способ разметки является причиной того, что во многих случаях, когда в кликсте реально есть только одно речевое событие, ему приписывается больше одной метки. В качестве примера можно привести эпизод из «Бриллиантовой руки», когда Геша (Андрей Миронов) сталкивает в воду мальчика с сачком (Максим Никулин) и произносит при этом: *А ну, щенок, в сторону! Пшёл отсюда!* Этот эпизод можно в равной степени описать как *требование*, *команда*, *распоряжение*—элементы всех трех речевых действий есть в этих двух фразах. Тем самым пользователю дается возможность получить этот эпизод по каждой из трех меток или по их совокупности.

На данный момент совокупность помет для аннотации речевых действий выглядит следующим образом (метки выстроены в группы для удобства ориентации; границы между этими группами условны и размыты, названия групп—в высшей степени условны, но такой способ подачи материала все же гораздо содержательней чисто алфавитного списка)—см. Таблица 6¹⁹.

¹⁹ В ряде случаев примеры не приводятся, поскольку в расписанных фильмах не нашлось по тем или иным причинам подходящего примера (например, проводы можно проиллюстрировать только фразами на том квазиязыке, на котором в «Бриллиантовой руке» общаются контрабандисты в исполнении Леонида Каневского и Григория Шпигеля). И, разумеется, все примеры приводятся в расчете на то, что фильм «Бриллиантовая рука» хорошо всем известен, так что в большинстве случаев за текстовыми примерами будут возникать воспоминания о соответствующих эпизодах фильма.

Таблица 6

Группы	Тип речевого действия	Примеры
Вопросы	вопрос общий*	
	вопрос частный*	
	вопрос косвенный*	<i>Ты что, с ума сошел?</i> <i>Может, к нам зайдем?</i>
	вопрос контактный*	<i>Слышь?</i>
	вопрос обратной связи*	<i>Ясно?</i> <i>Действительно?</i>
	вопрос нечленораздельный	[Горбунков:] <i>А? (показывает пальцем).</i> — [Милиционер Володя] <i>(понимает) А!</i> <i>Так надо.</i> — [Горбунков:] <i>Ага.</i>
Этикетные высказывания	переспрашивать**	<i>А вот я люблю песню про зайцев.</i> — Про кого? — <i>Про зайцев.</i>
	благодарность	<i>Разрешите поблагодарить вас, Николай Иванович, за вашу интересную, очень интересную экскурсию. Спасибо.</i>
	извинение	<i>Прости, друг!</i>
	пожалуйста (передача)	Пожалуйста <i>(отдает записку).</i> <i>Я вас очень буду ждать!</i>
	пожелание	Лучше бы я упал вместо тебя. — <i>Что ты, Геш, спасибо.</i>
	поздравление	
	предложить помощь**	<i>Разрешите, я вам помогу.</i>
	представиться**	Будем знакомы. Козодоев Геннадий Петрович. — <i>Очень приятно.</i>
	приветствие	<i>Доброе утро.</i> — <i>Здравствуйте.</i>
	приглашение	<i>Кто заказывал такси на Дубровку? — Я!</i> — Садитесь.
	проводы	
	прощание	<i>Ну, пора, турист!</i>
соболезнование		
спросить разрешения**	<i>Я щас (выходит из-за стола)</i> <i>Товарищ старший лейтенант, можно я вас при жене буду называть по званию?</i>	

Группы	Тип речевого действия	Примеры
Утверждения	аналогия	<i>Они будут следить за вами, а мы за ними. —Вроде живца? Понимаю, сам рыбак.</i>
	аргумент	<i>Врачи рекомендуют. Успокаивает нервную систему, расширяет сосуды.</i>
	вспомнить**	<i>Друг пригласил. —Да, кстати, как он себя чувствует после вчерашнего?</i>
	вывод	<i>Почему он интересуется? Что это, простое любопытство? Подозрительный тип.</i>
	догадаться**	<i>Ах! Рука! Его пытали! Как же я раньше не догадалась!</i>
	заявление	<i>Я знаю, что у тебя там! У тебя там не закрытый, а открытый перелом!</i>
	комментарий	<i>Это вам, сувенир Ага, следит!</i>
	констатация	<i>А может, домой? Так я и еду домой.</i>
	объявление	<i>Руссо туристо! Облико морале! Ферштейн?</i>
	объяснение	<i>Вообще, по правде говоря, я не хотел ехать. Я хотел купить жене шубу.</i>
	перечисление*	<i>А вы говорите —поскользнулся, упал, закрытый перелом, потерял сознание. Очнулся — гипс!</i>
	подсказка	<i>Кнопочку нажмите. —Да? (нажимает кнопку)</i>
	предположение	<i>Надеюсь, мы подружмся. —Конечно Наверно, мне бы надо... —Не надо.</i>
	предсказание	<i>[Геша:] Береги руку, Сеня, береги. —[Горбунков:] Надя расстроится. —[Геша:] Что делать!</i>
	предупреждение	<i>Но учтите —ровно в семнадцать ноль-ноль все должны быть на теплоходе.</i>
	рапорт	<i>Когда мы с ним таким образом познакомились, я изложил наш план.</i>
	рассказ	<i>Через несколько минут этот белоснежный красавец-лайнер отправится в очередной круиз, увозя в своих комфортабельных каютах большую группу советских туристов в увлекательнейшее путешествие.</i>

Группы	Тип речевого действия	Примеры
	сентенция	<i>Как говорит наш дорогой шеф, в нашем деле главное — этот самый реализм.</i>
	совет	<i>Товарищ капитан, а шо ж мне теперь делать? — А ничего! Отдыхайте, танцуйте, веселитесь.</i>
	сообщение	<i>Что с ним? — Действительно сильный вывих. Даже потерял сознание.</i>
	указание на кого/что	<i>О, вот он. Все в порядке. Товар, как в сейфе. Вон туда!</i>
	уговор	<i>Теперь договоримся о связи.</i>
	утверждение	<i>Эти кретины уверяли, что он был без сознания. Значит, этот лопух ничего не знает.</i>
	уточнение	<i>Сеня, ты уже дошел до кондиции? — До какой? — До нужной.</i>
Императивы	баюканье	<i>Спи, спи. — Сплю.</i>
	заказ	<i>Феденька, и хорошо бы дичь.</i>
	инструкция	<i>Фиш-стрит. Рыбна улица. Аптека Чиканук. — Лёлик, я всё прекрасно помню.</i>
	команда	<i>Начнем! К шефу!</i>
	настаивать**	<i>Пейте-пейте! Пейте!</i>
	поучение	
	предложение	<i>Выпейте. — Я не пью.</i>
	предостережение	<i>Осторожно! Осторожно! Береги руку, Сеня, береги!</i>
	приказ	<i>Докладывайте, что дальше. — Слушаюсь, товарищ полковник.</i>
	распоряжение	<i>Вам поручена эта операция, так что действуйте.</i>
	реклама	<i>О! Кто возьмет билетов пачку, тот получит... — Водокачку!</i>
	стоп!***	<i>Не надо! Иди! Брось эти шуточки!</i>
торопить**	<i>Надо действовать Цигель-цигель, ай-лю-лю!</i>	
требование	<i>Говори, что у тебя с рукой.</i>	

Группы	Тип речевого действия	Примеры
Модальные высказывания и перформативы	уговоры	
	успокаивать**	<i>Вот оно, началось. Только спокойно.</i>
	утешение	<i>Что делать! Такова селяви, как говорят у них.</i>
	беспокойство	<i>Не болит? — Неа.</i>
	ввод информации	<i>Когда мы с ним таким образом познакомились, я изложил наш план. Хотите верьте, хотите нет, а дело было так.</i>
	восклицание	<i>Теплоход через час уйдет! — Заткнись!</i>
	горе	<i>Шеф, все пропало, все пропало! Гипс снимают, клиент уезжает!</i>
	жалоба	<i>Ой! Ой боже мой!</i>
	клятва	<i>Но никому ни слова! — Клянусь!</i>
	молитва	
Шутливые/насмешливые высказывания	намерение	[Продавщица:] <i>К сожалению, нет.</i> — [Горбунков:] <i>Нет, да? Будем искать.</i>
	напоминание	<i>Семен Семеныч! А вещи?</i>
	ничего!***	<i>Ничего-ничего, у нас бывает и похуже.</i>
	обвинение	<i>А это? Элементы сладкой жизни!</i>
	обещание	<i>А про это (показывает) я сообщу куда следует. — Спасибо.</i>
	просьба	<i>Товарищ, товарищ, хорошо бы ай-лю-лю!</i>
	раскаяние	
	сочувствие	<i>Да... Бедняга.</i>
	угроза	<i>Лёлик, я не понимаю, о чем ты говоришь. — Сейчас поймешь.</i>
	уверенность	<i>Но я уверен, что до этого не дойдет.</i>
ирония	<i>Смешно, да? — (мрачно) Очень.</i>	
	намеки	<i>Товарищ, у вас когда самолет? — Да, пора.</i>
	насмешка	<i>Лопух! Такого возьмем без шума и пыли.</i>
	шутка	<i>Буду бить аккуратно, но сильно.</i>

Группы	Тип речевого действия	Примеры
Чужая речь	пересказ	<i>Поскользнулся... —Упал. —Упал. Закрытый перелом. Потерял сознание. Очнулся—гипс. —Правильно.</i>
	повтор подсказки	<i>Поскользнулся... —Упал. —Упал. Закрытый перелом. Потерял сознание.</i>
	цитирование	<i>Строго на север, порядка пятидесяти метров.</i>
Согласие	подтверждение	<i>Это каюта шестнадцать, или пардон, я ошибся? —Шестнадцатая.</i>
	понимание	<i>А что, если... —Не стоит. —Ясно.</i>
	признание	<i>Нет, я не трус... Но я боюсь.</i>
	разрешение	<i>Вы можете погулять по городу.</i>
	согласие подчиниться	<i>...как у вас там говорят, топай до хазы! —Хорошо, хорошо, я сейчас уйду... Одним словом, будьте больше на виду. —Ага.</i>
Отрицание	возражение	<i>Ну и вы тоже едете за границу первый раз? —Нет, я никуда не еду.</i>
	дистанцирование	<i>А я не знаю, как там в Лондоне, я не была. Может, там собака друг человека. А у нас —управдом друг человека.</i>
	запрет	<i>Только без рук! Никто не должен знать. —И Надя? —Никто.</i>
	недоверие	
	недовольство	<i>Ну что?</i>
	незнание	<i>Как же можно с человека срезать гипс незаметно? —Можно.</i>
	непонимание	<i>Наконец, с трупа. —Угу. С чьего трупа?</i>
опровержение	<i>Ну что же, все эти десять лет он пил, дебоширил и, так сказать, морально разлагался? —Ну нет. Вы знаете, все это время он искусно маскировался под порядочного человека.</i>	
отказ	<i>Ай-лю-лю потом. Нон, нит, нет, ни в коем случае.</i>	

Группы	Тип речевого действия	Примеры
	отрицание	А вы какие-нибудь сувениры с собой берете? — Нет-нет-нет-нет-нет.
	поправка	Ну-ка, пошевели пальчиками. Нет, не этими, вот этими.
	сомнение	Хороший человек. — Вы думаете?
	спор	
	удивление	Какая лекция?
Апеллятивы	звать**	Геша! Геша!
	обращение	Леди, синьора, фрау, мисс! К сожалению, ничего не выйдет!
	обращение к животному	Кс-кс!
	отклик	Сеня! — А? —А ты Софи Лорен видел? — Неа. Геннадий Петрович! — Да?
	пароль	Черт побери! Черт побери! Шьорт побьери! Добрый вечер, Борис Савельич. Я заказал Феде дичь . Очень прошу вас.
	привлечение внимания	Товарищи, внимание! Сейчас у вас свободное время. SOS!
	призыв	Остановитесь! Возьмем! —Не положено.
	призыв к порядку	Максим! Ты что делаешь!
Мелиоративные высказывания	похвала	Упал. Закрытый перелом. Очнулся —гипс. — Правильно!
	похвальба	На полагающуюся мне по закону премию я, по совету друзей, решил приобрести автомашину «Москвич». Новая модель!
	тост	За твою премию. —Дай бог, не последняя.
	удовлетворение	Алё! В девять часов вас устраивает? —Ага. — Ну и хорошо.
Пейоративные высказывания	брань	Идиот! Дитям мороженое!
	оскорбление	Виктор Николаевич! Вы что, читать не умеете, а?

Группы	Тип речевого действия	Примеры
	критическое замечание	<i>Не умеешь ты врать, Сеня.</i>
	порицание	<i>Ну разве можно так! Извинись щас же перед дядей!</i>
	проклятье	<i>Шоб ты издох! Шоб я видел тебя в гробу в белых тапках!</i>
	стыдить**	<i>Какой позор!</i>
	упрек	<i>Что ж ты меня бросил-то?</i>

* Одной звездочкой обозначены речевые действия, для которых принято лингвистическое обозначение;

** двумя звездочками – речевые действия, для обозначения которых не нашлось соответствующего существительного, поэтому используется глагол;

*** тремя звездочками – речевые действия, которые не отражены в языке и для обозначения которых используется наиболее частое словесное их воплощение.

Здесь следует сделать небольшое отступление. На материале английского языка, а точнее, на материале английских корпусов с размеченными речевыми действиями, уже ведутся исследования, которые позволяют до некоторой степени автоматизировать разметку речевых действий (см. об этом [Webb et al. 2008]). Логика здесь следующая. На базе корпуса, в котором речевые действия размечены вручную, вычлениются ключевые слова и словосочетания, частотность которых в тех или иных речевых действиях существенно выше, чем их частотность в целом по аннотированному корпусу. Если такие слова и словосочетания обнаруживаются, то принимается, что их можно расценивать как показатели того или иного речевого акта. Так, например, словосочетание *can you* считается одним из показателей общего вопроса, словосочетания *where is* или *which is* — показателями частного вопроса (для русского языка, например, словосочетание *разрешите поблагодарить* может определенно расцениваться как показатель благодарности, *пшел* — как показатель приказа, команды, распоряжения, требования, и т.д.). Мы назвали только очевидные случаи — ясно, что при сплошном исследовании аннотированного мультимедийного корпуса на этот предмет проявятся гораздо менее очевидные, но столь же несомненные результаты. Отсюда — возможность использовать эти

ключевые слова и словосочетания для разметки речевых действий в большом корпусе, в частности, в НКРЯ.

Следующее поле метаразметки кликстов—*полнота* речевого действия (см. Таблица 7).

Таблица 7

Степень полноты	Примеры
Полное	Большая часть примеров
Автопрерывание	<i>Говори, что у тебя с рукой.—Я и говорю—шел по улице, поскользнулся, уп<ал>... Ну, вот.</i>
Вопрос без ответа	<i>А я взял.—Водку? (смеются)</i>
Наложение реплик	<i>Извините, что так поздно. Сами знаете...—Ну что вы!—...общественное дело прежде всего.</i>
Незаконченное речевое действие	<i>Значит, вы еще никогда...—Конечно, не был. Мы вообще дальше Дубровки никуда не ездили.</i>
Прерванное	<i>(обращается к отсутствующему Геше) Ты зна<ешь>... (видит вместо Геси осла, испуганно) А!</i>
Продолженное	<i>[Михаил Иванович:] Если вы нам понадобится...—[Горбунков:] Вы ко мне приедете.—[Михаил Иванович:] Если мы вам будем нужны...—[Горбунков:] Я вызываю такси на свое имя.—[Михаил Иванович:] Приеду я или мой товарищ.</i>
Жест вместо слова	<i>Вы уж извините, что я... (показывает руку)</i>

Далее, к параметрам метаразметки относится *манера говорения*. Здесь выделяются такие типы: норма (большая часть примеров), быстрая речь (например, речь персонажа Рины Зеленой в фильме «Подкидыш»), голос за кадром, декламация, дефекты дикции (например, речь персонажа Ролана Быкова в фильме «По семейным обстоятельствам»), диктовка, дубляж (например, женский голос, который дублирует разговоры персонажей Леонида Каневского и Григория Шпигеля в «Бриллиантовой руке»), крик, напевающее произнесение, невнятная речь, оговорка (*А у вас нет такого же, но с берла... перламутровыми пуговицами?*), пение, говорение «про себя», пьяный разговор, произнесение «на слезе» или «на смехе», скандирование, «чрево вещание» (т.е. говорение с максимально неподвижной мимикой, когда говорящий старается, чтобы его речь

не была услышана и замечена кем бы то ни было, кроме адресата), чтение, шепот.

Следующий признак метаразметки—*типы повторов*, имеющих место в том или ином кликсте (см. Таблица 8).

Таблица 8

Тип повтора	Пояснение	Пример
Многократный		<i>Черт побери! Черт побери! Черт побери!</i>
Однократный		<i>А что, что я должна была подумать?</i>
Неоднословный		<i>Что с вами? Что с вами?</i>
Однословный		<i>«Михаил Светлов?»—Да-да.</i>
Передразнивание	Повтор реплики собеседника или ее части с издевательской интонацией	<i>Но тот тоже сказал пароль «черт побери»!—Черт побери, черт побери!</i>
Переспрос		<i>Вы не знаете, зачем Володька усы сбрил?—Усы?</i>
Повтор с интенсификатором	Повтор слова с включением интенсификаторов (очень, никогда, всегда, исключительно, абсолютно, никакой и проч.)	<i>Разрешите поблагодарить вас ... за вашу интересную, очень интересную экскурсию</i>
Повтор с разной интонацией		<i>(недоуменно) Черт побери... (находит решение, радостно) Черт побери!</i>
Смена адресата	Продолжение реплики адресуется новому слушателю	<i>[Ладыженский:] (Горбункову) Ну, будете у нас на Колыме... (Козодоев закашлялся) (Козодоеву) ...будете у нас на Колыме—милости просим!</i>
Эхо	Повтор со сменой говорящих и с сохранением типа речевого действия	<i>[Первый контрабандист:] Пароль старый, черт побери?—[Второй контрабандист:] (подтверждает) Черт побери.—[Первый контрабандист:] (эхо) Черт побери...</i>

Далее, размечается **несловесное наполнение** кликстов (вокальные жесты, междометия, физиологические действия). Здесь предварительно выделены следующие типы (проиллюстрированы по естественным причинам в тексте статьи могут быть только некоторые из них): аккомпанемент-да (*Минуточку! Связь будем держать так...—Угу.—...если вы нам понадобитсяь...*), аккомпанемент-ну (*Давайте грузить! (грузят пьяного в коляску мотоцикла).—Ну!—Ну!—Ну!*), аккомпанемент-э (*Козодоев пытается на крыльце дома обойти Михаила Ивановича Э... э... э... (падает, встает) Пардон!*), боль²⁰, вздох, волнение, восхищение, втянуть носом воздух, выдох, звонок (*дзынь, бип*), дразнить (лаять, щелкать зубами, показывать язык, сопровождая это соответствующими звуками...), жалость, заполнитель паузы (эканье, меканье и проч.), зевок, изобразительное (изображение звуками теплохода, шампанского и проч.), испуг, кашель, насмешка, недоверие, недовольство, недоумение, неожиданность, обращение (к кошке, птице и проч.), одобрение, опа!, отклик (*Сеня!—А?*), отрицание, плач, плевков, понимание (*Гражданочка! (женщина отшатывается) А...*), поцелуй, пренебрежение, привлечение внимания, призыв к тишине, приниматься, припоминание, причмокивание, свист, смешок, согласие, спохватиться, угроза, удивление, физическое напряжение, хмыканье, холодно, цокать языком, шмыгнуть носом.

И наконец, предусмотрены поля **количество говорящих, пол говорящих** (мужской, женский, смешанный—т.е. говорят и мужчины, и женщины²¹) и **язык**, на котором говорят (здесь по умолчанию предполагается русский, однако может быть указано, что для говорящего характерен акцент; если в кликте говорят на иностранных языках, то это указывается; среди языков предусматриваются квазиязык, как в «Бриллиантовой руке», и тайный язык, как в «Друг мой, Колька!»).

²⁰ То есть вокальные жесты и междометия, связанные с испытываемой болью, испугом и т.д.

²¹ Следует отметить, что многие дети и подростки, играющие в кино, «озвучиваются» женщинами, и определить это на слух может только профессионал высокого класса. Поэтому если пользователя интересуют ситуации озвучания ребенка или подростка мужского пола женщиной, то он должен будет специальным образом находить соответствующие кликсты посредством обращения к полю «возраст персонажа» в зоне разметки жестов, см. ниже.

4.

Перейдем к изложению системы метаразметки, принятой для описания жестового наполнения кликстов/клипов (подчеркнем, что практически все нижеизложенное базируется на работах [Григорьева и др., 2001], [Крейдлин 2004]).

Разметка жеста включает следующие основные поля:

1. Социологическая характеристика
2. Аксессуары
3. Кратность жеста
4. Основной орган
5. Активный орган
6. Характеристика жеста
7. Полнота жеста
8. Аутентичность жеста
9. Эмоциональное сопровождение жеста

1. Социологическая характеристика дается жесту по следующим параметрам.

1.1. *Имя* говорящего (актера), если таковое известно или может быть упомянуто публично без этических и юридических ограничений.

1.2. *Пол персонажа* (если таковой известен—например, в мультфильмах иногда трудно определить пол того или иного персонажа).

1.3. *Пол говорящего* (для кино- и мультфильмов). Здесь, помимо стандартного мужского и женского пола предусматриваются: *актер, играющий женщину* (например, Георгий Милляр в роли Бабы-Яги, Олег Табаков в роли няни в «Мэри Поппинс, до свиданья!»), *актер, притворяющийся женщиной* (например, Александр Калягин в роли тетушки Чарли в «Здравствуйте, я ваша тетя!»), *актриса, играющая мужчину* (чисто теоретический и умозрительный вариант), и *актриса, притворяющаяся мужчиной* (например, Лариса Голубкина в «Гусарской балладе»). Все эти варианты существенны для исследования гендерных аспектов жестикуляции.

1.4. *Возраст говорящего (актера)*, если таковой известен. Возраст дается описательно (напомним, что точный возраст говорящего (актера) прописан в социологической разметке устного

корпуса и, соответственно, будет легко доступен и пользователю мурко)—*ребенок, подросток, взрослый, пожилой*.

1.5. **Возраст персонажа** (для кино- и мультфильмов). Возраст также дается описательно. Несоответствие возраста актера и возраста персонажа может быть интересно для возрастных характеристик жеста. При этом, очевидно, для некоторых героев мультфильмов возраст персонажа определить довольно затруднительно—например, для Винни-Пуха, Чебурашки, Крокодила Гены или для Карлсона, про которого известно лишь, что он мужчина в самом расцвете лет.

1.6. **Социальная ситуация**, отраженная в клипе. Этот параметр полностью совпадает с аналогичным параметром для разметки речевых действий (см. выше), поэтому размечаться должен лишь для клипов, т.е. в отсутствие речевой составляющей.

2. **Аксессуары**. Аксессуарами считаются все предметы, так или иначе задействованные в жестикуляции. Это могут быть *удлинители* (например, карандаш, который покусывают при раздумьях вместо пальца), *спойлеры* (предметы, которые мешают осуществлению данного жеста в полном объеме,—например, когда человек разводит одной рукой, а не двумя, осуществляя жест «развести руками», поскольку в одной из рук у него находится трость). Кроме того, это могут быть предметы, нейтральные по отношению к жестикуляции, например, сигарета, зажатая между пальцев, которая не влияет на протекание жеста, или, напротив, предметы, являющиеся центром данного жеста, например, та же сигарета в ситуации прикуривания. Разметка аксессуаров позволит не только «укрупнить» описание жестов (т.е. «развести руками» при наличии спойлера нет смысла описывать как жест «развести рукой»—это по-прежнему жест «развести руками», но «испорченный» наличием постороннего предмета), но и осуществлять некоторые культурологические и психологические наблюдения—например, можно будет узнать, когда первый раз на советском экране появилась курящая женщина, или, например, как связаны с проявлениями тех или иных эмоций разные манеры закуривать или держать сигарету в руке.

3. **Кратность жеста**. Жесты делятся на *однократные* (отделенные от других жестов наличием своей собственной экспозиции и фазы затухания, иначе говоря, наличием семантической паузы

до и после жеста) и *многократные*—имеющие общую экспозицию и фазу затухания, а также амплитуду, в среднем укороченную по сравнению со сходным однократным жестом. При подборе названия для жеста (см. ниже, п. 6) однократные жесты описываются глаголами совершенного вида, а многократные—несовершенного (таким образом, например, однократный кивок будет обозначен глаголом *кивнуть*, а многократное кивание—глаголом *кивать*).

4. **Основной орган**—часть тела человека, в зоне которой жест осуществляется. Выделены следующие основные органы: *голова, корпус, рука, руки, нога, ноги*.

5. **Активный орган**—движущаяся, активная (инициативная) часть основного органа, которая, собственно, и формирует жест. Распределение здесь следующее.

Таблица 9

Основной орган	Активные органы
Голова	<i>брови, глаз, глаза, голова, губы, зубы верхние, лицо, подбородок, рот, язык</i>
Корпус	<i>корпус, плечи, спина</i>
Рука	<i>кисть, мизинец, палец большой, палец указательный, палец указательный + палец большой, палец средний, палец указательный + палец средний, пальцы, рука</i>
Руки	<i>кисти, пальцы, пальцы указательные, руки</i>
Нога	<i>стопа, голень</i>
Ноги	<i>ноги</i>

6. Характеристика жеста состоит из трех разделов—название, значение и тип жеста. **Название жесту** присваивается 1) либо наиболее стандартным его общезыковым обозначением (чаще всего с помощью глагола), 2) либо условным обозначением (например, *стоп!* или *двинуть ладонь к собеседнику*). **Значение жеста**—это функция, которую жест выполняет в данном конкретном случае его употребления. **Тип жеста**—условная группа однотипных по значению жестов. Предварительно нами вычленено 13 типов жестов. В таблице 10 им в соответствии поставлены лишь наиболее простые и частотные примеры.

Таблица 10

Тип жеста	Значение жеста	Название жеста
Жесты внутреннего состояния	испуг задуматься готовность	отпрянуть взяться за подбородок поправить галстук
Дейктические жесты	общеуказательный самоидентификация фиксация объекта	показать пальцем показать подбородком демонстрация показать на себя пристукнуть по чему-л.
Декоративные жесты	общедекоративный	поправить одежду поправить прическу подтянуть брюки откинуть голову
Изобразительные жесты	действие (играть на трубе) объект (вода) качество (точность)	перебирать пальцами перебирать пальцами соединить большой и указательный пальцы
Корпоративные жесты	молитва воинская субординация	поклониться, стоя на коленях отдать честь
Пейоративные жесты	подумаешь! дурак! передразнивание	вскинуть руку сплюнуть мотать головой
Поисковые жесты	оценивать обстановку оценка времени поиск слова	оглядываться посмотреть на часы трясти рукой
Регулирующие жесты	привлекать внимание призыв к порядку иди!	махать руками посмотреть строго выдвинуть подбородок
Жесты—речевые действия	угроза отрицание согласие	грозить кулаком качнуть головой кивнуть
Риторические жесты	предвосхищение согласия фиксация объекта	кивнуть пристукнуть по чему-л.

Условные жесты	похвала	показать большой палец
	отказ	показать кукиш
	тост	чокнуться
Физиологические жесты	больно	тереть больное место
	горько	сморщиться
Этикетные жесты	прощание	махать рукой
	извинение	прижать руку к груди

Это лишь очень небольшая часть расписанного материала, но она дает представление о вычлененных крупных группах жестов.

7. *Полнота жеста*. В стандартном случае жест является полным, т.е. протекает в полном объеме, с естественной аккомодацией и «гаплогией» экспозиций и фаз затуханий жестов, которые соседствуют в жестовом синтаксисе при реальном, а не искусственном функционировании жестового языка. Однако изредка встречаются типичные случаи неполного осуществления жеста, которые разумно отмечать при аннотировании клипов. Предварительно вычленяются следующие разновидности явлений такого рода: *автопрерывание* (жестикулирующий добровольно прекращает данную жестикуляцию), *прерывание* (жест принудительно прерывается собеседником или обстоятельствами), *трансформация* (один жест по ходу осуществления превращается в другой), *редукция* (жестикулирующий лишь намекает на возможность осуществления какого-либо жеста или максимально редуцирует его—например, вместо полного жеста «поднять палец вверх», когда вертикально вверх поднимается не только указательный палец, но и вся кисть, жестикулирующий лишь слегка отрывает от горизонтальной поверхности указательный палец, а кисть остается лежать на поверхности).

8. *Аутентичность жеста*. В нормальном случае жест является аутентичным, т.е. отражает внутреннее состояние и речевые намерения самого жестикулирующего. Однако изредка встречаются случаи неаутентичности, которые также следует по возможности отмечать. Эти случаи таковы: *зеркальный жест*—жестикулирующий повторяет жест за собеседником, *жест, показанный на себе*—жестикулирующий на себе показывает жесты тех, кто в данный момент является персонажем или адресатом его речи, *притворный*

жест—используется в ситуации, когда жестикулирующий очевидным образом неискренен и делает притворные жесты, например, сочувствия или радости за собеседника. Кроме того, для кино есть смысл отдельно вычленять т.н. случаи *игровых жестов*, которые часто сопровождают вставные музыкальные номера в фильмах.

9. И, наконец, под эмоциональным сопровождением жеста имеется в виду, сопровождается ли жест улыбкой, смехом или плачем.

Представляется, что даже такая, довольно огрубленная и местами неточная разметка жестов в мультимедийном корпусе будет иметь большое значение для исследователей. Вероятные и даже неизбежные недостатки этого аннотирования наверняка будут компенсироваться возможностью обращаться к жесту не только напрямую, но и через «словесный» поиск, а также поиск «от речевого действия». Таким образом, все эти типы разметок будут подстраховывать друг друга и в целом позволят максимально снизить уровень шума при поиске информации в мурко.

5.

В завершение следует сказать о том, что, конечно, большое значение для мурко будет иметь единообразие в работе будущих разметчиков. Практика работы над нкря показала, что даже в такой «точной» части разметки, как морфологическая, есть немалое количество точек, в которых разметчиками (как людьми, так и автоматическими парсерами) могут быть приняты *разные* решения, у каждого из которых есть свои плюсы и свои минусы (наиболее очевидный случай—разметка видовых пар, где могут быть приняты два альтернативных решения: признавать видовую пару манифестацией одного глагола или парой разных и независимых глаголов). Еще больший разброс возможностей имеет семантическая разметка. Надо ли говорить, что разметка жестов и речевых действий имеет максимальную тенденцию к «размытости» результатов. Для любого корпуса это—одна из самых больших опасностей, поскольку неустойчивость разметки затрудняет работу пользователя с корпусом, что в значительной степени обесмысливает существование последнего. И здесь создатель корпуса должен стремиться не столько даже к *правильности* принятого решения, сколько к его *единообразию*.

Грубо говоря, если при разметке какого-то одного явления принято неточное, если не сказать неправильное решение, то именно это решение, а никакое другое, должно быть принято и для всех остальных аналогичных явлений. В этом случае пользователь, будучи не согласным с содержательной стороной работы создателей корпуса, будет, однако, понимать, какие именно ему нужно совершить действия, чтобы найти данное неправильно или неточно размеченное явление (не говоря уже о том, что при единообразном *неправильном* решении есть хорошие шансы заменить его на единообразное *правильное* решение, тогда как при разное решений и чересполосице правильностей и неправильностей правка такого рода представляет иногда чрезвычайно сложную лингво-программистскую задачу).

Именно поэтому к созданию мурко в полном объеме, т.е. с дополнительной разметкой, описанной выше, нет смысла приступать, не имея специального «рабочего места разметчика», которое будет общаться с разметчиком в диалоговом режиме и в значительной степени «вести» его от речевого действия к речевому действию и от жеста к жесту по единообразно устроенным «тропинкам», привязанным к тем или иным относительно точным реперным точкам. Содержательная часть рабочего места на данный момент уже разработана автором настоящей статьи, на программистском уровне поставленная задача в настоящее время в целом решена Михаилом Кудиновым (МГУ им. М. В. Ломоносова). Для иллюстрации приведем одну такую «тропинку» для разметки одной из групп жестов.

- Разметчику предлагается выбор—**однократный/многократный жест**
 - Если выбран однократный жест, ему предлагается выбор **основной орган** *голова, корпус, рука, руки, нога, ноги.*
 - Если выбран основной орган *голова*, предлагается выбор **активный орган** *брови, глаз, глаза, голова, губы, зубы верхние, лицо, подбородок, рот, язык*
 - Если выбран активный орган *голова*, предлагается выбор **пассивный орган** *кисть, плечо, нет пассивного органа.*
 - Если выбран пассивный орган, то предлагаются на выбор следующие жесты:

Значение жеста	Название жеста	Тип жеста
усталость	опереться головой на руку	физиологический жест
скука	опереться головой на руку	жест внутреннего состояния
огорчение	опереться головой на руку	жест внутреннего состояния
нервозность	опереться головой на руку	жест внутреннего состояния
дистанцирование	опереться головой на руку	жест внутреннего состояния

Таким образом, даже если пользователь вдруг совершенно не согласится с нашей трактовкой жеста *опереться головой на руку*, он будет точно знать, в каких клипах такой жест найдется,—ему надо будет при запросе обратиться к параметрам ‘основной орган=голова’, ‘активный орган=голова’, ‘пассивный орган=кисть’. Ну и, разумеется, пользователь получит этот жест, обратившись к значениям жеста—скука, усталость, огорчение, нервозность, дистанцирование—и к типам жестов—внутреннего состояния или физиологическим.

6.

Итак, по предварительным планам, мурко будет состоять из двух частей.

1. Кликсты, связанные с соответствующими скриптами (скрипты при этом будут размечены так, как принято в нкря, т.е. будут иметь соответствующую метаразметку, морфологическую, семантическую и социологическую разметку). Для подготовки этой части корпуса достаточно разрезать исходные фильмы и звуковые файлы на кликсты и сопоставить каждому кликсту соответствующий участок скрипта.
2. Кликсты/клипы, получившие дополнительную метаразметку с точки зрения имеющих в них речевых действий и жестов. Понятно, что на создание второй части мурко потребуется затратить гораздо больше усилий, чем на первую (хотя и к первой части относиться легкомысленно было бы неразумно, поскольку она предполагает значительный объем ручной работы по «нарезке» исходного материала, которую заведомо невозможно автоматизировать). Поэтому, вероятно, часть корпуса с дополнительной

разметкой должна быть как-то специально отобрана, в частности, очевидным образом, она должна быть сбалансирована хронологически, чтобы позволить производить те или иные диахронические изыскания.

При этом совершенно ясно, что обе части мурко позволяют ставить и решать совершенно новый класс лингвистических задач, степень разнообразия которых на данном, подготовительном этапе трудно даже оценить. И, разумеется, значение мурко далеко выходит за пределы собственно лингвистики—корпус позволит обращаться к общефилологическим, культурологическим, историческим, психологическим, педагогическим и инженерным проблемам, не только о решении, но и о постановке которых сейчас, в отсутствие национального мультимедийного корпуса, нельзя и думать.

ЛИТЕРАТУРА

- Григорьева и др. 2001—Григорьева С. А., Григорьев Н. В., Крейдлин Г. Е. Словарь языка русских жестов. М.—Вена: 2001
- Гришина 2007—Е. А. Гришина. О маркерах разговорной речи (предварительное исследование подкорпуса кино в Национальном корпусе русского языка) // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007» (Бекасово, 30 мая—3 июня 2007 г.). С. 147–156
- Гришина, Савчук 2008—Гришина Е. А., Савчук С. О. Корпус звучащей русской речи в составе Национального корпуса русского языка. Проект // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2008» (Бекасово, 4–8 июня 2008 г.). С. 125–132
- Капанадзе 1986—Капанадзе Л. А. Разговорная речь и киноязык // Л. А. Капанадзе. Голоса и смыслы. Избранные работы по русскому языку. М.: 2005. С. 228–231.
- Крейдлин 2004—Крейдлин Г. Е. Невербальная семиотика. М.: 2004.
- Шаронов 2008—Шаронов И. А. К вопросу о разграничении эмоциональных междометий и модальных частиц // Компьютерная лингвистика и интеллектуальные технологии. Труды меж-

- дународной конференции «Диалог 2008» (Бекасово, 4–8 июня 2008 г.). С. 569–573
- Blache et al. 2008—Philippe Blache et al. Creating and exploiting multimodal annotated corpora // [LREC'2008]
- Brutti et al. 2008—A. Brutti et al. WOZ Acoustic Data Collection for Interactive TV // [LREC'2008]
- Busso, Narayanan 2008—Carlos Busso and Shrikanth S. Narayanan. Recording audio-visual emotional databases from actors: a closer look // [LREC'2008]
- Cullen et al. 2008—Charlie Cullen et al. Emotional Speech Corpus Construction, Annotation and Distribution // [LREC'2008]
- Devillers, Martin 2008—L. Devillers, J-C. Martin. Coding Emotional Events in Audiovisual Corpora // [LREC'2008]
- Fék et al. 2008—Márk Fék et al. Multimodal Spontaneous Expressive Speech Corpus for Hungarian // [LREC'2008]
- Forbes-Riley et al. 2008—Kate Forbes-Riley et al. Uncertainty Corpus: Resource to Study User Affect in Complex Spoken Dialogue Systems // [LREC'2008]
- Gallo et al. 2008—Carlos Gómez Gallo et al. Production In A Multimodal Corpus: How Speakers Communicate Complex Actions // [LREC'2008]
- Geertzen et al. 2008—Jeroen Geertzen et al. Evaluating Dialogue Act Tagging with Naive and Expert Annotators
- Georgila et al. 2008—Kallirroi Georgila et al. A Fully Annotated Corpus for Studying the Effect of Cognitive Ageing on Users' Interactions with Spoken Dialogue Systems // [LREC'2008]
- Gnjatović, Rösner 2008—Milan Gnjatović, Dietmar Rösner. On the Role of the NIMITEK Corpus in Developing an Emotion Adaptive Spoken Dialogue System // [LREC'2008]
- Grishina 2007—E. Grishina. Text Navigators in Spoken Russian. // Proceedings of the workshop “Representation of Semantic Structure of Spoken Speech” (CAEPIA'2007, Spain, 2007, 12–16.11.07, Salamanca), Salamanca, 2007. P. 39–50
- Hennoste et al. 2008—Tiit Hennoste et al. From Human Communication to Intelligent User Interfaces: Corpora of Spoken Estonian // [LREC'2008]
- Knight, Tennent 2008—Knight, D., Tennent, P. Introducing DRS (The

- Digital Replay System): A tool for the future of Corpus Linguistic research and analysis // [LREC'2008]
- Kostoulas et al. 2008—Theodoros Kostoulas et al. A Real-World Emotional Speech Corpus for Modern Greek // [LREC'2008]
- LREC'2008—<http://www.lrec-conf.org/proceedings/lrec2008>
- Marasek, Gubrynowicz 2008 —Krzysztof Marasek, Ryszard Gubrynowicz. Design and Data Collection for Spoken Polish Dialogs Database // [LREC'2008]
- Möller et al. 2008—Sebastian Möller et al. Corpus Analysis of Spoken Smart-Home Interactions with Older Users // [LREC'2008]
- Nallasamy et al. 2008—Udhyakumar Nallasamy et al. NineOneOne: Recognizing and Classifying Speech for Handling Minority Language Emergency Calls // [LREC'2008]
- Sainz et al. 2008—Iñaki Sainz et al. Subjective evaluation of an emotional speech database for Basque // [LREC'2008]
- Savino et al. 2008—Michelina Savino et al. Integrating Audio and Visual Information for Modelling Communicative Behaviours Perceived as Different // [LREC'2008]
- Stoia et al. 2008—Laura Stoia et al. SCARE: A Situated Corpus with Annotated Referring Expressions // [LREC'2008]
- Strauß et al. 2008—Petra-Maria Strauß et al. The PIT Corpus Of German Multi-Party Dialogues // [LREC'2008]
- van Son et al. 2008—R.J.J.H. van Son et al. The IFADV corpus: A free dialog video corpus // [LREC'2008]
- Webb et al. 2008—Nick Webb et al. Cross-Domain Dialogue Act Tagging // [LREC'2008]
- Wilks et al. 2008—Yorick Wilks et al. Dialogue, Speech and Images: The Companions Project Data Set // [LREC'2008]
- Wilson 2008—Theresa Wilson. Annotating Subjective Content in Meetings // [LREC'2008]