Итогом выполнения проекта должны стать:

– формат данных (на основе XML-TEI), адаптированный для описания древних текстов, рукописей и их фрагментов; при этом необходимо решить такие проблемы, как представление пересекающихся фрагментов в разметке XML, описание дат в неявном виде (например, первая половина XI века) и другие;

– средства загрузки документов в ИПС «Манускрипт» для последующей работы с ними, а также возможность соединения с уже описанными фрагментами, организованными в иерархии и в некоторых случаях связанными со словарями;

– возможность поиска по текстам, фрагментам, а затем и внутри фрагментов;

– инструменты редактирования текстов (фрагментов, представленных в указанном формате с возможностью сохранения);

– средства выгрузки документов.

Выполненная работа позволит объединить усилия нескольких коллективов для более активного и глубокого исследования рукописных памятников славянской культуры.

Размеченный корпус диалогов как ресурс моделирования диалога: организация и разметка Эстонского корпуса диалогов[1]

О. Герасименко, Т. Хенносте, М. Койт, Р. Кастерпалу,
А. Ряэбис, К. Страндсон, М. Вальдисоо, Э. Вутть
Тартуский университет, Эстония

The Estonian Dialogue Corpus is collected with the aim of developing the dialogue system using the natural language. The spoken dialogues (884 dialogues, 155000 running words) are used to study the rules and norms of the human-human communication; the corpus also includes human-computer dialogues (21, 2500 running words) collected by the Wizard of Oz method used to study the role behaviour of the users and information provider. The presentation considers the means and levels of transcription and annotation dialogues and also the application of the corpus.

Эстонский корпус диалогов создан лингвистами и компьютерными технологами Тартуского университета с целью исследовать речевое взаимодействие в естественных диалогах и моделировать общение между человеком и диалоговой системой, которая должна следовать нормам человеческой коммуникации.

1. Состав корпуса

Основная часть корпуса состоит из естественных устных диалогов (758 справочных телефонных диалогов и 106 непосредственных диалогов, всего 884 диалога объемом в 155000 слов).

Устные диалоги сохраняются в цифровом или оцифрованном формате .wav, а в текстовом виде представлены в транскрипции Джефферсон [Jefferson 2004], которая следует принципам анализа речевого взаимодействия (conversation analysis). Транскрипция фиксирует явления, позволяющие проследить динамическое построение диалога из реплик и интонационных единиц: движение

[1] Работу поддерживает Эстонский научный фонд (грант 5685).

интонации, ударность слов, замедление и убыстрение темпа речи, длину пауз, наложение двух реплик и др.

Малая, но важная часть корпуса — 22 письменных диалога объемом в 2500 слов — представляет собой компьютерные симуляции по методу «волшебника Оз». 11 пользователей тестировали программу, предоставляющую транспортную информацию на естественном языке, но в действительности роль программы исполнял человек. Диалоги представляют собой логи бесед пользователя и «системы».

2. Уровни аннотации

2.1. Морфология

Часть устных диалогов корпуса (20000 слов) размечена морфологически. Автоматическая разметка осуществлена с помощью морфологического анализатора Estmorf, разработанного для литературного эстонского языка; омонимия снята вручную (два разметчика и экспертное обсуждение, которому сопутствовало исследование проблем снятия морфологической омонимии в устной эстонской речи в сравнении с письменной).

2.2. Речевые акты

Ни одна из рассмотренных схем разметки не отвечала целям моделирования и исследования естественных диалогов с точки зрения речевого взаимодействия, поэтому нам пришлось создать свою типологию разметки, в которой мы попытались объединить положительные черты других типологий [Gerassimenko et al. 2004].

Типология состоит из 126 диалоговых актов и основывается на рассмотрении диалога как социального взаимодействия. Речевые акты подразделяются на две группы — акты, составляющие смежные пары (например, вопросы и ответы), и одиночные акты (например, побудительные реплики обратной связи). С другой стороны, акты подразделяются на относящиеся к предмету разговора (например, информационные акты) или к управлению диалогом (например, инициирование исправлений).

Восемьсот двадцать устных диалогов и двадцать один диалог ВОЗ размечены двумя аннотаторами, разногласия которых устранены в ходе экспертного обсуждения. Точность разметки, рассчитанная для 45 устных диалогов разного типа, составляет 0.74. Ре-

чевые акты размечаются вручную при помощи вспомогательной программы.

Пример 1 (естественный диалог в переводе на русский).
((звонок)) | RIE: ЗВОНОК |
V: `справочная | RIJ: ОТВЕТ НА ЗВОНОК | | RY: ИДЕНТИФИКАЦИЯ |
`Криста | RY: ИДЕНТИФИКАЦИЯ |
здравствуйте? | RIE: ПРИВЕТСТВИЕ |
(.)
H: здравствуйте, | RIJ: ПРИВЕТСТВИЕ |
=я `хотел бы ((название фирмы)) (.) < дис`петчера > или `какой-нибудь `но- номер. | DIE: ДИРЕКТИВ |
(0.8)
V: м минуточку. | DIJ: ОТСРОЧКА |
(5.5)
V: телефон об`служивания [то есть,] да? | KYE: ОБЩИЙ ВОПРОС | | VTE: УТОЧНЕНИЕ УСЛОВИЙ ОТВЕТА |
H: [да.] | KYJ: ДА | | VTJ: УТОЧНЕНИЕ УСЛОВИЙ ОТВЕТА |
V: по неподтвержденным данным номер < четыре четыре семь? (0.5) девять восемь? (.) [один] ноль. > | DIJ: ИНФОРМАЦИЯ |
H: [.hh] | YA: ИНОЕ |
(.)
H: спасибо. | RIE: СПАСИБО |
V: пожалуйста? | RIJ: ПОЖАЛУЙСТА |

2.3. Коммуникативные стратегии

Мы исходим из понятия коммуникативной стратегии и конструктивной модели диалога (Constructive Dialogue Model, CDM) [Jokinen 1995]. Коммуникативная стратегия используется участником диалога для построения следующей реплики как реакции на предыдущую реплику партнера. Коммуникативные стратегии, таким образом, выражают когерентность диалога так же, как смежные пары диалогических актов, но на более обобщенном уровне.

Для определения коммуникативных стратегий в CDM используются четыре контекстуальных фактора:

– ожидаемость реплики;
– связь реплики с темой;
– достигнутость целей говорящего;
– инициатива говорящего или партнера.

Все контекстуальные факторы бинарны, соответственно, коммуникативных стратегий 2*4=16.

Наиболее удачной структурой представления коммуникативных целей является магазин целей (stock) и принцип «последним вошел, первым вышел»: первый вопрос клиента порождает главную цель, помещаемую на «дно» магазина, последующие уточняющие вопросы порождают новые цели, помещаемые последовательно одна над другой. По мере достижения верхних целей они удаляются; все цели диалога считаются достигнутыми, когда магазин пуст.

Коммуникативные стратегии размечены в 20 естественных и 20 симуляционных диалогах.

Пример 2 (симуляционный диалог в переводе на русский язык).

Реплика	Стратегия	Магазин
Клиент: *Как доехать до Пярну из Тарту до полудня?*	ожидаем.-связн.-достигн.-говорящ.-*начало/конец*	–
Компьютер: *Автобус выходит в 5 утра. Автобус выходит в 8 утра.*	ожидаем.-связн.-недостигн.-говорящ.-*продолжение темы*	Тарту-Пярну
Интересует ли вас время прибытия? ...	неожидаем.-связн.-достигн.-говорящ.-*новый диалог*	Тарту-Пярну

Размеченный корпус используется в теоретических и прикладных исследованиях (напр., [Hennoste et al. 2005]) и в работе над автоматическим распознанием речевых актов (с помощью ограничений и нейронных сетей [Fishel 2004]), а также в построении первых диалоговых систем на эстонском языке. Эти системы, Транспортный и Театральный Агент, общаются с пользователем преимущественно по ключевым словам (в письменном и устном регистре) и с применением знаний о структуре информационных диалогов и поддиалогов.

Список литературы

Mark Fishel 2005. Dialogue Act Recognition in Estonian Dialogues Using Artificial Neural Networks. In: Proceedings of the Second Baltic Conference on Human Language Technologies, Tallinn, 4–5 April 2005, 249–254.

Olga Gerassimenko, Tiit Hennoste, Mare Koit, Andriela Rääbis, Krista Strandson, Maret Valdisoo, Evely Vutt. Annotated Dialogue Corpus as a Language Resource: An Experience of Building the Estonian Dialogue Corpus. The First Baltic Conference “Human Language Technologies. The Baltic Perspective”. Commission of the Official Language at the Chancellery of the President of Latvia, Riga, 2004, 150–155.

Tiit Hennoste, Olga Gerassimenko, Riina Kasterpalu, Mare Koit, Andriela Rääbis, Krista Strandson, Maret Valdisoo. Questions in Estonian Information Dialogues: Form and Functions. Text, Speech and Dialogue. 6th International Conference TSD 2005. Springer, 2005, 420–427.

Gail Jefferson 2004. Glossary of transcript symbols with an introduction. In Lerner, G.H. (Ed). *Conversation Analysis: Studies from the first generation*. Amsterdam/Philadelphia: John Benjamins, 13–31.

Kristiina Jokinen. 1995. Rationality in Constructive Dialogue Management; URL: http://cl.aist-nara.ac.jp/lab/papers/kris/aaai.ps (used 20.05.2006).

Приложение

1. Транскрипционные знаки

спад интонации	.
полуспад интонации	,
подъем интонации	?
короткая пауза (макс. 0.2 с.)	(.)
длина паузы в секундах	(2.0)
наложение	[text]
слияние независимых единиц	=
растянутый звук	::
ударное слово	`
прерванное слово	do-
вдох	.hhh
убыстрение темпа	> text <
замедление темпа	< text >
сомнительный отрезок	{text}
неразборчивый отрезок	{---}

2. Типология речевых актов

I Акты, составляющие смежные пары
1.1 Акты управления диалогом
1.1.1 Коммуникация
1.1.1.1 Ритуалы (RIE RIJ)
1.1.1.2 Смена темы
1.1.2 Разрешение проблем

1.1.2.1 Исправление, инициированное партнером
1.1.2.2 Проверка контакта
1.1.2.3 Уточнение условий ответа (VTE VTJ)
1.2 Информационные акты
1.2.1 Директивы (DIE DIJ)
1.2.2 Вопросы (KYE KYJ)
KYE: общий вопрос
KYE: общий вопрос, ожидающий развернутого ответа
KYE: альтернативный вопрос
KYE: специальный вопрос
KYE: иное
KYJ: да
KYJ: нет
KYJ: согласное нет
KYJ: иной ответ на общий вопрос
KYJ: альтернатива: одна
KYJ: альтернатива: обе
KYJ: альтернатива: третий выбор
KYJ: альтернатива: отрицание
KYJ: альтернатива: иное
KYJ: развернутый ответ
KYJ: отсутствие информации
KYJ: отказ
KYJ: иное
1.2.3 Мнение
II Одиночные акты
2.1 Акты управления диалогом
2.1.1 Коммуникация
2.1.1.1 Ритуалы (RY)
2.1.1.2 Обратная связь
2.1.2 Разрешение проблем
2.1.2.1 Исправление
2.2 Информационные акты
2.2.1 Основные акты (YA)
2.2.2 Дополнения основных актов (пояснение, уточнение)

Примечание: Подробно расписана группа вопросов и ответов — наиболее частотных речевых актов; эстонская аббревиатура, предшествующая названию акта, содержит информацию о группе и о позиции акта в смежной паре.

Корпус древнерусских агиографических текстов СКАТ: современное состояние и перспективы развития

А. С. Герд, И. В. Азарова, Е. Л. Алексеева, Е. С. Иванова
Санкт-Петербургский государственный университет, Россия

The Corpus of Russian hagiographic texts of the 16–17th centuries at present comprises 52 texts or 500 000 word-tokens, represented in 2 formats: as text files and Microsoft Word files; the texts are provided with a word form index. 10 texts have been published; they are available to Internet users in the PDF and XML formats. The work is under way to provide all texts with the morphological information. Automatic normalization of varying Church-Slavonic spelling is another important task.

Корпус агиографических церковнославянских текстов XVI–XVII вв. на кафедре математической лингвистики Санкт-Петербургского государственного университета начал создаваться в конце 70-х годов. Работа началась с создания картотеки житий святых русской церкви, похвальных слов, сказаний, в которой учитывались исследования и издания этих текстов; были изысканы средства для образования фонда фото- и ксерокопий рукописей житий, находящихся в разных рукописных хранилищах Петербурга, который постоянно пополняется. Тогда же, в конце 70-х, началась работа по вводу текстов житий в компьютер. К настоящему времени корпус охватывает 52 жития, их общий объем — более 500 тыс. словоупотреблений.

Параллельно формированию базы данных было начато изучение грамматики, словообразования конкретных текстов. В результате к концу 1996 г. вышло в свет три обобщающие книги, которые содержат систематическое описание именного склонения, глагольного спряжения и именного словообразования памятников русской агиографической литературы XVI в., опубликован ряд