А. В. Костыркин

Корпус японской разговорной речи[1]

настоящей статье описывается крупнейший японский проект по созданию корпуса устной речи, его история, состав, виды и принципы языковой разметки, в частности разные виды транскрипции, сегментация, морфологическая, синтаксическая, дискурсивная и фонетическая разметка.

1. История проекта

Корпус спроектирован и создан японским Государственным институтом национального языка (ГИНЯ) совместно с Лабораторией по исследованиям в области телекоммуникаций и Токийским технологическим университетом. Основная работа по сбору и обработке материала выполнена в ГИНЯ. Куратор проекта — проф. Фуруи Садаоки из Токийского технологического университета [Maekawa

[1] Исследование выполнено при частичной финансовой поддержке Российского фонда фундаментальных исследований, грант № 07-06-00061. Автор благодарит научного сотрудника Гос. ин-та национального языка Маруяма Такэхико за помощь, оказанную при подготовке статьи.

2008]. Проект осуществлен в период с 1999 по 2003 г. и на момент завершения стал самым большим размеченным корпусом устной речи в мире [Uchimoto 2003].

Главная цель проекта – подготовить в достаточном объеме качественные лингвистические данные, на основе которых будут разрабатываться системы автоматического распознавания речи, в частности, будет происходить обучение статистических систем. Была поставлена задача получить представительный корпус современного общенационального языка (gendai kyootsuu nihongo), поэтому диалекты, устаревший язык и случаи смешения японской речи с иностранной в него не включались.

Вышло два издания корпуса — первое (июнь 2004 г.) и второе — исправленное и дополненное (май 2008 г.). Сейчас корпус используется в Японии и за ее пределами в более чем 280 проектах. Японское название корпуса – Nihongo hanashikotoba koopasu; официальное английское название — The Corpus of Spontaneous Japanese.

Создание корпуса активно освещалось в периодических научных и популярных изданиях [Maekawa et al. 2001]. Основным документом, описывающим принципы создания корпуса, основные виды разметки и ее лингвистические критерии является свободно распространяемый отчет «Метод построения корпуса разговорной японской речи» [Nihongo 2006].

2. Объем и состав корпуса

Записан 661 час спонтанной речи, что соответствует около 7,52 млн. слов. Запись осуществлялась в формате Digital Audio Tape с использованием конденсаторных микрофонов. Записи прорежены до 16kHz, 16 бит. Растекстовка, т.е. перевод аудиозаписей в текстовую форму, и дальнейшее транскрибирование проводилась в соответствии со специально выработанной системой записи в двух формах: смешанным иероглифико-азбучным письмом и только азбучным. Для записанных таким образом текстов проведена морфологическая разметка, выполненная в двух формах: в терминах так называемых коротких и длинных морфологических единиц (см. ниже). В корпусе выделена часть общей протяженностью 44 часа (около 500 тыс. слов), называемая «ядром», для которой осуществлена также фонетическая, просодическая, дискурсивная и синтаксическая разметка.

Около 90% записей корпуса составляет монологическая речь, остальные 10% – диалогическая речь, чтение вслух письменного текста, чтение вслух транскрипции устного текста. Всего записана речь 1417 человек. Следующая таблица дает представление о составе корпуса.

Виды речи	**Кол-во говорящих**	**Кол-во файлов**	**Монолог / диалог**	**Спонтанная речь / чтение текста**	**Продолжительность, часов**
Академическая публичная речь	819	987	монолог	спонтанная	274,4
Искусственные выступления	594	1715	монолог	спонтанная	329,9
Прочие выступления	*16	19	диалог	спонтанная	24,1
Искусственные выступления интервьюируемых	*16	16	монолог	спонтанная	3,4
Интервью на темы научных докладов	*10	10	диалог	спонтанная	2,1
Диалог на заданную тему	*16	16	диалог	спонтанная	3,1
Свободный диалог	*16	16	диалог	спонтанная	3,6
Чтение текста вслух	*248	507	монолог	чтение	15,5
Повторное чтение вслух	*16	16	монолог	чтение	5,5
Итого	1417	3303			661,6

Таблица 1. Виды и объем записей в составе корпуса
*Говорящие входят в число тех, кто участвовал в записи первых двух типов выступлений

Под *академической публичной речью* имеются в виду научные доклады, которые записывались вживую в течение трех лет в трех научных обществах. Продолжительность большинства из этих записей—от 12 до 25 мин., имеются также записи протяженностью более 1 часа. Так называемые *искусственные выступления* включают записи речи на заданную тему перед аудиторией из 3-5 человек. Запись происходила в раскрепощенной обстановке. Большей части говорящих предлагалось по 3 темы довольно широкого содержания, продолжительность записи по каждой теме составила в среднем 12 мин. Темы сообщались участникам за 48 часов до записи, при этом им запрещалось готовить письменный текст выступления, но

рекомендовалось продумать простой план речи. Ста говорящим было позволено выбрать для своего выступления одну-две темы по своему желанию. Вот примеры предлагавшихся тем: «Что было в вашей жизни радостного или приятного», «Что было в вашей жизни печального или тяжелого», «Расскажите о городе или районе, где вы живете», «Дайте объективное объяснение чему-то, что вы хорошо знаете или чем вы интересуетесь», «Что больше всего запомнилось в жизни», «Газетные, журнальные новости последних лет», «Три предмета, которые возьмете с собой на необитаемый остров», «Как сделать (что-либо), как приготовить (что-либо)», «История (чего-либо)», «Что, кто вам больше всего дорого/дорог?», «Что бы мне хотелось сделать для 21-го века и чего бы не хотелось». Записи академической речи и выступлений сделаны на одном и том же оборудовании, но часть из них произведена не в студии, а в обычном помещении, поэтому есть различия в акустике. Под *чтением текста вслух* имеется в виду запись чтения одним участником двух коротких отрывков из распространенных книг по естественным наукам. Продолжительность записи 3–4 мин. *Искусственные выступления интервьюируемых* записывались с теми же говорящими, которые участвовали в записи диалогической речи. *Интервью на темы научных докладов* записывались после соответствующих научных докладов и искусственных выступлений, описанных выше. Продолжительность каждого интервью составляет около 10-15 мин. Под *повторным чтением вслух* имеется в виду запись транскрибированного текста научного доклада тем же говорящим. При этом в чтении воспроизводились паузы и запинки, как они представлены в транскрипции. Все такие записи осуществлялись в звукоизолированном помещении. В случае диалогической речи каждый из двух говорящих находился в отдельной комнате с общим окном, их речь записывалась на разные каналы.

Для сохранения анонимности участников из текстов корпуса удалены все имена, которые как-либо указывают на личности говорящих.

3. Разметка

Главным принципом при разметке корпуса было стремление избежать односторонних решений в случаях неоднозначной атрибуции речевых единиц. Там, где существующие описания японского

языка не дают четких критериев определения парадигматических и/или синтагматических границ речевой единицы или же ее частеречной принадлежности, авторы старались учесть все разумные интерпретации, отразить их в разметке и указать на наиболее вероятную из них.

Разметка корпуса, по мнению его авторов, должна эксплицировать наблюдаемые проблемы, а не скрывать их. Выявленные, но пока не решенные проблемы авторы считают одной из важных составляющих корпуса и ценным материалом для будущих исследователей и составителей корпусов устной речи.

3.1. Транскрипция

Записанная речь разбита на отдельные *транскрипционные единицы*, которым сопоставлена разметка. Границы единиц проводятся в речи там, где есть пауза более 200 мс, либо пауза более 50 мс после лексической или грамматической формы, способной завершать предложение (финитная форма глагола, заключительная частица типа *wa, ne, yo, ka*, приветствие, частицы со значением «да», «нет»). Если определение звуковых границ единицы неоднозначно, выбираются более широкие границы. Если звук губ, сопровождающий артикуляцию, не удается отделить от речи, он также включается в состав единицы. Если фонация единицы начинается с взрывной или аффрикаты, то в качестве начала единицы устанавливается момент 50 мс до взрыва. Не транскрибируются и далее не учитываются обособленные одно-двухморные короткие звуки или продолжительные тихие звуки, которые невозможно интерпретировать.

Выделены транскрипционные единицы 4-х типов:

A. Вербальные единицы.
B. Голосовые звуки, издаваемые говорящим (смех, плач, кашель, звуки, сопровождающие произнесение заполнителей типа *ши*, *аа*, *апоо*).
C. Все остальные звуки
D. Ошибки, возникающие при чтении вслух (только для повторного прочтения текстов).

Единицы типов A и B не могут пересекаться во времени, но могут образовывать вложения типа *et-* <кашель> *-to*. Единицы типа C мо-

гут пересекаться с единицами других типов, исключение составляют случаи типа короткого кашля во время продолжительного смеха, которые остаются без учета.

Каждой записанной единице присвоен индивидуальный номер, указано время начала и конца фонации, номер канала, на который записан звук, а также собственно лингвистическая разметка. Границы единиц, а также тип невербальных единиц (шум, кашель и т. д.) определялись автоматически. Текст транскрипции и разметка вводились вручную путем прослушивания звука и наблюдения на экране компьютера формы звуковой волны и спектрограммы. Разметка единиц четырех названных выше типов заносилась в четыре отдельные поля.

Лингвистическая разметка состоит из двух частей: правой и левой. Справа дается так называемая базовая транскрипция, слева — фонетическая.

Базовая транскрипция	**Фонетическая транскрипция**
今までの	& イママデノ
人生で	& ジンセーデ
一番	& イチバン
印象深かった	& インショーブカカッタ
こと	& コト

Рис. 1. Пример пяти последовательных текстовых единиц, для которых даны параллельно два вида транскрипции (отделены друг от друга амперсандом).

3.1.1. Базовая транскрипция

Базовая транскрипция использует иероглифико-алфавитную запись и предназначена для максимально простого чтения текста, а также для текстового поиска. Для этой цели при помощи строгих правил из записи исключена вариативность и неоднозначность, используется строго ограниченное множество иероглифов, знаков и сочетаний слоговой азбуки, числовых знаков и знаков препинания. Это достигается за счет разработанных правил унификации различных случаев варьирования, которыми изобилует японская письменность. Перечислим некоторые такие правила:

1. Многие японские слова допускают несколько вариантов записи — либо одними иероглифами, либо одной *каной*[2], либо иероглифами в сочетании с каной.

Если у слова одинаково употребительны и азбучная, и иероглифическая форма записи, то выбирается последняя. Это позволяет при последующем автоматическом морфологическом анализе получать более точные результаты. Азбучная форма выбирается, если она является устоявшейся (решение принималось на основе обследования материала газет и лексико-иероглифических справочников, выпущенных компанией NHK). Если использование азбучной и иероглифической формы связано с выражением разных лексико-грамматических значений, то полнозначный вариант записывается с иероглификой, а вариант со служебным значением одной каной (ср. 上げる ‘поднимать’ vs. あげるбенефактив от ‘поднимать’). Если слово используется как словообразовательный компонент, то в составе сложного слова оно дается по возможности в той же форме (ср. 掛ける и 追い掛ける).

2. Для служебных слов は, へ, を принята историческая запись (は передается как есть, а не как わ), для сложных слов со второй основой на つ *tsu*, в том числе слов с редупликацией, принята морфологическая запись, и дается づ, а не ず (ср. つづら).

3. Долгота гласных *a*, *i*, *u* передается их повтором, долгота гласной *e* — либо ее повтором, либо добавлением *i*, гласной *o* — либо ее повтором, либо добавленим *u*.

4. При нормализации *окуриганы* — вариативного написания каной определенных частей слов — для слов изменяемых частей речи выбирается графически наиболее длинный вариант (行なう, а не 行う). Для слов неизменяемых частей речи выбирается вариант с каной (買い値, а не 買値). При этом сделано исключение для ряда слов, для которых написание без каны стало устоявшимся (например, используется запись 取締役, а не 取り締まり役).

5. Для случаев лексической омонимии, которые на письме противопоставлены графически, используются разные формы записи: 表わす / 現わす. Для трудно разграничимых случаев полисемии выбирается графический вариант, передающий наиболее общее зна-

[2] Каной называется слоговая японская азбука — хирагана или катакана.

чение (для глагола 逢う ‘встречаться (о любовниках)’ выбирается запись 会う ‘встречаться’). Если такое обобщение затруднительно, то противопоставление сохраняется, как, например, в случае слов 意志 *ishi* ‘воля’ и 意思 *ishi* ‘намерение’.

Использование разных видов графических единиц — иероглифики, хираганы, катаканы, цифр и букв английского алфавита в целом регулируется следующими правилами:

1. Иероглифика используется помимо записи слов *ваго* и *канго*[3] для случаев автонимного употребления отдельных иероглифов и обрывков слов, когда граница обрыва совпадает с морфологической, и такой обрывок может быть записан одним иероглифом. Допускается использование иероглифики первого и второго уровней, определенных стандартом JIS X 0208–1990. Если у иероглифа есть устаревший и новый графический варианты, то используется современный вариант или тот, который есть в JIS первого уровня (например, из пары вариантов 証 и 證 выбирается первый).

2. Хирагана, помимо установившейся записи слов ваго и канго, используется для идеофонов[4], заполнителей (*etto* ‘ну’), названий букв и звуков (*kana no a* ‘знак «а» каны’), обрывков слов, когда граница разрыва не совпадает с морфологической (*ryo…ryoohoo no* «об…, обоих»). Все допустимые сочетания знаков азбуки заданы конечным списком, в котором сочетания для записи дифтонгов разделены на основные и периферийные. К основным относятся все палатализованные слоги с гласными *a*, *u* и *o*, используемые при записи канго и ваго и стандартно выделяемые в японских учебниках. К периферийным отнесены слоги с лабиализованными согласными (クワ *kwa*), с гласными *e* и *i* (シェ *she*, ミェ *mie*), слоги на *f-* (ファ *fa*, フィ *fi*) и другие, которые встречаются в заимствованных словах.

3. Катакана используется для записи заимствованных слов, неизменяемых частей слов, образованных от заимствованных (таких как *toraburu* «доставлять неудобства» от англ. *trouble*), иностранных имен собственных, кроме китайских и корейских, терминов и жар-

---

[3] Ваго—исконно японская лексика; канго—заимствованная китайская лексика.

[4] Конкретные идеофоны, как правило, могут быть отнесены к японскому или китайскому лексическим классам, но записывают их крайне непоследовательно—то хираганой, то катаканой.

гонизмов с устоявшейся катаканной записью, названий растений и животных, а также для названий букв и звуков применительно к катаканным словам. Обрывки слов и оговорки в заимствованных словах также записываются катаканой.

4. Слова, которые записываются буквами английского алфавита, обязательно сопровождаются их транслитерацией, записанной катаканой согласно определенным правилам.

5. Числа записываются цифрами и вместе с этим полностью расписываются словами по-японски.

6. Допускается использование ограниченного числа знаков препинания. Японские знаки препинания существуют двух видов — полноширинные и полуширинные, в данном корпусе используются только полноширинные знаки.

Знак повтора иероглифа 々 может при необходимости использоваться многократно, например 点々々, но не используется в случаях повтора иероглифов в сокращениях сложных слов, таких как 自自公[5]. Срединная точка « ・ » используется для отделения имени или инициала имени от фамилии в иностранных именах, для разделения членов сочиненных именных групп, а также для снятия неоднозначности в именных сцеплениях[6]. Для того, чтобы скрыть некоторые имена собственные (в целях сохранения анонимности записей. см. выше), используется знак *batsu* « × ». Знак *maru* «○» служит только для записи нуля в японском тексте.

Составители корпуса постарались отразить в транскрипции отличие случайных отклонений от произносительной нормы от регулярных случаев такого рода. Случайные отклонения, такие как эмфатическое удлинение гласных и согласных, помечаются как ошибочные, но транскрибируются при этом нормализованными формами. Случаи же регулярных, уже закрепившихся в языке фонетико-морфологических модификаций, характерных для устной речи, транскрибируются как есть, т.е. считаются утвердившимися единицами разговорного языка. На основе системного анализа

---

[5] Сокращение названий трех партий: 自民党 «Либерально-демократическая», 自由党 «Либеральная», 公明党 «Партия чистой политики».

[6] Например, в сочетании 哲学・教育的背景 «философское и педагогическое образование» точка поставлена, чтобы не было прочтения 哲学教育の背景 «философско-педагогическое образование».

корпусного материала был выделен перечень таких разговорных выражений. Вот некоторые их них:

1. Выпадение слога на *r-* в конце глагольной формы в сочетаниях с отрицанием *nai*, субстантиватором *n/no* или связкой *da*: *shira nai* → *shinnai* 'не знаю', *aru n da* → *anda* 'имеется', *suru daroo* → *sundaroo* 'вероятно, сделает'.

2. Выпадение гласной *i* во вспомогательном глаголе *iru* после деепричастной формы на *te/de*: *mite iru* → *miteru* 'смотрит', *yatte oite* → *yattoite* 'сделай'

3. Модификация формы на *-ba*: *kakeba ii* → *kakya ii* 'лучше написать', *mi nakereba ii* → *mi nakya ii* 'можно не смотреть'.

Во всех этих случаях наблюдается тяготение аналитической формы к синтетической.

4. Редукция конечного гласного в субстантиваторе *mono*: *ima kita mon de* 'только что пришел'.

5. Редукция конечного гласного в слове *nani* в значении «почему», «какой»: *nande*, *nanto*, *nanda*.

В процессе разметки происходило составление двух словарей, в которых регистрировались все единицы, встречающиеся в тексте. Первый словарь предназначен для человека, в нем для каждой единицы регистрировалась ее правильная и возможная ошибочная разметка, с тем чтобы избежать подобных ошибок в дальнейшем. К концу проекта объем словаря составил 110 тыс. единиц. Во втором словаре запоминались все сочетания каны и соответствующей ей иероглифики. Учитывались также вероятные ошибки, которые может допустить человек при вводе определенных слов. Эти сведения использовались потом для повышения точности автоматического преобразования в иероглифику каны, вводимой человеком. Эта процедура осуществлялась при помощи программы Kanna.

3.1.2. Фонетическая транскрипция

Каждой записи в базовой транскрипции сопоставлена транскрипция фонетическая. Она выполнена катаканой и предназначена для поиска по произношению и для снятия неоднозначности иероглифики в базовой транскрипции. Она также может быть использована при исследовании фонетических и фонологических вариаций,

возникающих в спонтанной речи. Приняты следующие правила транскрибирования:

1. Используется только катакана. В целях унификации записи знаки ряда *t-* ヂ*ji* и ヅ *zu* исключены из использования и заменены знаками ряда *s-* ジ *ji* и ズ *zu* соответственно (слово チヂム, например, записывается вопреки общим правилам как チジム).

2. Служебные слова は, を и へ записываются как произносятся, т.е. как ワ, オ и エ соответственно.

3. Искаженные слова, произнесенные небрежно или с ошибкой, записываются как есть, но сопровождаются восстановленным по контексту предположительно правильным произношением.

4. Случаи нечеткого разграничения между долгой гласной и дифтонгом передаются на выбор либо знаком долготы, либо повтором гласной, либо двумя гласными, если между двумя гласными нет морфемной или другой границы (カーサン *kaa-san* ‘мама’, ケイロ *kei-ro* ‘маршрут’). При этом глагольные формы типа *yaroo* ‘давай сделаем’ считаются одной морфемой. Если же морфемная граница между гласными есть, то допускается только повтор гласной (ダイイチ *dai-ichi* ‘номер один’).

5. Удлинение гласных и согласных, которое не фиксируется нормализующими словарями (*sugooi* вместо *sugoi*, *tottemo* вместо *totemo*), маркируются специальными тегами.

Выделены 4 случая неоднозначно произнесенных слов:

1. Произношение неоднозначно, но можно понять, что это за слово. В этом случае произносительный вариант помечается как ошибочный, при нем дается правильный.

2. Произношение неоднозначно, и нет уверенности, существует слово или нет. Решение о транскрипции принимается на основе контекста.

3. Произношение неоднозначно, и слово определить невозможно. В транскрипции со знаком вопроса перечисляются все вероятные варианты интерпретации.

4. Неоднозначность вызвана тем, что у слова более одного варианта произношения. В этом случае на основе словарей и частотных списков, которые строятся на основе корпуса, устанавливается вариант по умолчанию и он выбирается в качестве транскрипции. Альтернативные варианты также фиксируются, но со знаком вопроса.

3.2. Система тегов

Разработана система тегов и правила разметки ими различных вербальных и невербальных звуков. Для каждого из четырех типов единиц A–D, названных выше в разделе 3.1, определен свой набор тегов. В таблице 2 приведены примеры тегов, которые используются для разметки вербальных единиц.

Таблица 2. Примеры тегов для вербальных единиц

Тег	**Область использования**	**Пример помеченных тегом единиц**
(D), (D2)*	Слово, разделенное на фрагменты. Тег D2 используется только для исправлений служебных слов, состоящих из одной моры.	(D こ) これ これ(D2 は)が
(W)	Оговорка, искаженное, ослабленное произношение. В скобках слева от точки с запятой выражение, которое трактуется как ошибочное, справа – его исправление.	(Wミダリ; ヒダリ)**
(?)	Нет уверенности в правильности понимания, определении лексической единицы, либо в выборе иероглифической записи имеется несколько вариантов.	(? タオングー) (? あのー、あんのー)
(F)	Заполнитель пауз (filler), эмоциональное междометие	(F あの), (F うわっ)
(M)	Метаязыковое выражение (автонимное употребление, цитация)	(M わ) は (M は) と表記する («ва» пишется как «ха»)
(O)	Иностранные слова (не заимствования), устаревшие слова, диалектизмы — все, что не является основным предметом описания в данном проекте.	(O ザッツファイン)
(A)	Слова, в словарной форме которых используются знаки помимо иероглифов и азбуки. Используется, в частности, для иероглифической записи чисел.	(A イーユー; EU) (A 百十九; 119)番

(K)	По какой-либо причине, например, из-за вставки заполнителя пауз, стало невозможным иероглифическое выражение единицы.	(K たち(F んー) ばな;橘)
(S)	Разговорное выражение, не зарегистрированное в транскрипционном словаре	(S ほりゃ)
(B)	Из-за неграмотности говорящий допустил ошибку в прочтении иероглифической записи. Сюда входят, в частности, смешение китайских и японских чтений иероглифов, ошибочно пропущенные озвончение, назализация или геминация на стыке морфем.	脱力 & (B ダツリキ;ダツリョク) 夢見話 & ユメミ(B ハナシ;バナシ) 悪化 & (B アクカ;アッカ) 何だって & (B ナニ;ナン)ダッテ
(笑) (泣) (咳) (あくび)	Отмечают случаи, когда невербальные звуки (смех, плач, кашель, зевание) накладываются по времени на вербальные.	(笑 ナニソレ)
(L)	Шепот, бормотание, другие случаи понижения голоса	(L アノコレナンダッケ)
<H>	Произвольное удлинение гласных	ソレデ<H> 私 & ワタシ<H>
<Q>	Произвольное удлинение согласных	カイ<Q>セキ
<FV>	Неопределенный гласный звук	ソレデ<FV>
<P>	Пауза длиной более 200 мс внутри краткой единицы (см. ниже)	オ<P:00453.373-00454.013>モイ

Отдельно выделяются теги для невербальных звуков, см. Таблицу 3

Таблица 3. Примеры тегов для невербальных единиц.

<息> <笑> <泣> <咳>	Шум дыхания, смех, плач, кашель (не совпадающие по времени с речью)
<ベル>	Звонок во время доклада
<拍手>	Хлопки аудитории
<雑音>	Любой другой вид шума

Для каждого типа тегов определено, может ли он использоваться в базовой или фонетической транскрипции, а также какие множества символов он может содержать. Для неоднозначных случаев постановки тегов разработаны детальные правила, учитывающие контекст, частеречные классы слов, между которыми возможен выбор, их морфологический состав. Так, если контекст не позволяет однозначно определить, является ли данная единица *sono* заполнителем ‘как бы’ или прилагательным ‘тот’, она помечается как заполнитель, которому приписана альтернативная интерпретация. Таким образом, пользователю корпуса дается возможность найти единицу по любой из интерпретаций и уточнить ее характеристику самостоятельно. Возможны случаи вложенной записи тегов, когда обозначаемые ими единицы частично совпадают во времени.

3.3. Диалоговая разметка

В случае диалогов каждый из двух говорящих записан на один из двух каналов, обозначаемых в разметке L и R. На всех интервью ведущий записан на канал L. Поскольку голоса собеседников не накладываются друг на друга, разметка таких текстов велась так же, как и монологической речи. Диалоги записаны в форме одного текста, реплики даны в порядке их произнесения. Если говорящие друг за другом по частям произносят одно слово, то обе части слова помечаются тегом D как фрагменты.

3.4. Разметка текстов, прочитанных вслух

Часть выступлений после перевода их из звуковой в текстовую форму была прочитана вслух, при этом в каждом случае текст читал человек, который произносил его изначально. Читающего просили озвучивать все записанные оговорки и их исправления, заполнители, паузы и т.п. Поскольку состав читаемого текста известен заранее, в разметке таких текстов нет тегов <FV> или (?), обозначающих нераспознанные единицы. Случаи расхождения речи с текстом, такие как оговорки и возвраты назад по тексту, помечались специальным тегом <朗読間違い> «ошибка воспроизведения». Части, добавленные говорящим к тексту от себя, помечены тегом <X>.

3.5. Деление текста на бунсэцу

В корпусе не выделяется такой текстовой единицы, как предложение (см. ниже). Максимальной по протяженности размечаемой единицей является синтагма-бунсэцу. Это традиционно выделяемая в японской грамматике единица, состоящая, как правило, из сочетания полнозначного слова с цепочкой относящихся к нему примыкающих служебных слов. Тексты корпуса сегментированы на такие синтагмы, и это, во-первых, облегчает разметку текста, в частности, соотнесение базовой транскрипции с фонетической, во-вторых, позволяет использовать полученные единицы в дальнейшем синтаксическом и дискурсивном анализе.

В растекстовке бунсэцу отделены друг от друга знаком новой строки, специальных тегов, указывающих на их границы, не предусмотрено. Если внутри одного бунсэцу оказывается пауза длиннее 200 мс, то оно разбивается и записывается в две или более строки. Поскольку такие случаи немногочисленны, можно считать, что в целом одна строка в растекстовке корпуса соответствует одному бунсэцу. Вот основные случаи, в которых проводятся границы бунсэцу:

1. После цепочки служебных слов и вспомогательных глаголов.
2. После подлежащего и тематической группы.
3. После определительных групп, как изменяемых (renyou), так и неизменяемых (rentai).
4. После глагола в срединной или финитной форме, а также в форме императива.
5. Справа и слева от наречий.
6. После междометий.
7. После имен без оформляющих их послелогов (dokuritsu-kaku).
8. Внутри именного сцепления (taigen-renzoku), если у части сцепления есть собственное определение.
9. Между аппозитивными членами.
10. Между сочиненными членами.

Даже если названные правила требуют выделения границ бунсэцу, этого не происходит в следующих исключительных случаях: между именами и фамилиями, внутри сложных слов идиоматического характера, внутри составных географических названий, названий праздников, товаров, сложных названий растений, названий теле-

передач, музыкальных и художественных произведений, математических формул и в некоторых других специально оговоренных случаях.

Из этих исключений есть свои исключения, которые распространяются на случаи, характерные именно для разговорной речи, такие как полные или частичные исправления уже сказанного, прерывание фразы на середине (iisashi), вставка одной фразы внутрь другой, в частности не допустимое в письменной речи разделение знаменательного и подчиняющего его служебного слова финитными глаголами[7], которые могут быть пояснениями или привлекающими внимание слушателя оборотами.

3.6. Морфологическая разметка

Морфологическая разметка состояла из выделения собственно морфологических единиц и определения их лексико-грамматической интерпретации [Ogura 2008]. 1 млн. слов был размечен вручную, на что ушло более 2-х лет. Остальные 6,5 млн. слов размечено автоматически [Uchimoto 2003]. При этом ручная разметка для повышения эффективности работ частично была автоматизирована.

Для текстовых форм по специальным критериям определяются следующие признаки:

1. *Текстовые границы словоформы*. Проблема определения границ слов (go), ровно как и определения самого понятия слова, в японской лингвистике до сих пор не имеет окончательного решения [Gengo 2006]. Обследования текстов, которые проводились в ГИНЯ до создания корпуса устной речи, не дали универсального решения: в каждом исследовании в зависимости от его целей в качестве минимальной единицы лексического описания приходилось выбирать текстовые единицы разной протяженности.

Поскольку удовлетворить всем потребностям пользователей корпуса заведомо невозможно, авторы поставили две максимально общих задачи — во-первых, дать возможность исследовать лексику и грамматику разговорного языка, во-вторых, позволить выявлять лингвистические особенности именно устной речи. Между этими целями есть противоречие. С одной стороны, для исследования

[7] Случаи типа 弁別率 | ですね | を «степень различия | COP+PART | ACC», где между именем и показателем прямого дополнения вставлена связка.

лексического состава корпуса желательно выделить минимальные текстовые единицы[8]. С другой стороны, членение на минимальные по протяженности единицы исключает из получаемого в результате лексикона единицы, которые характерны именно для устной речи. Чтобы преодолеть это противоречие, в корпусе проведено разделение текстовых единиц на два вида — долгие и краткие — и морфологическая разметка сделана для единиц обоих видов. Долгие единицы соответствуют бунсэцу (см. выше). Большинство долгих единиц составляют сложные существительные (*kokuritsu-kokugo-kenkyujo* «Гос. институт национального языка») и глаголы (*tabe-akiru* ‘пресытиться’). К их числу относятся также устойчивые сочетания двух служебных слов (*de-wa* ‘итак’), а также служебных слов с глаголами (*ni+yoru+te = niyotte* ‘посредством чего’).

Краткие единицы — это минимальные единицы, имеющие в современном языке значение. Выделяются шесть классов кратких единиц: ваго, канго, гайрайго, символы, имена людей, топонимы. В некоторых случаях краткие единицы состоят из двух минимальных единиц, под которыми понимаются морфемы или просто словообразовательные элементы, записываемые одним иероглифом. Краткими единицами считаются заполнители пауз и обрывки слов. Отдельные правила предусмотрены для выделения кратких единиц внутри слитно произнесенных слов (yuugoo) и сокращений.

Всего в корпусе выделено 7,52 млн. кратких и 6,31 млн. долгих единиц.

2. *Словарная форма слова*. За счет того, что у каждой лексемы в устном корпусе помимо косвенных форм имеется множество сокращенных, плохо артикулированных или ненормативных текстовых реализаций, число и вариативность форм одной лексемы в устном корпусе выше, чем в письменном. Для обеспечения полноты и точности корпусного поиска для всех текстовых форм определя-

[8] Авторы осознают опасность чрезмерного дробления, которое может привести к появлению шума при текстовом поиске. Например, слово *itarutokoro* ‘везде, всюду’ не должно находиться при поиске по слову *itaru* ‘идти, достигать’. В свою очередь желательно иметь возможность найти терминологические сочетания типа *gengo-shigeki* «языковой стимул», *gengo-moderu* «языковая модель» и как целостные единицы, и как единицы, в составе которых есть слово *gengo* «язык».

ется их словарная форма, по которой проводится разграничение между различными лексемами. Словарная форма состоит из двух частей — из азбучной (daihyoo-kei) и азбучно-иероглифической записи (daihyoo-hyooki). Азбучная запись отражает чтение слова и его морфологический состав (например, для 或いは ‘или же’ чтение записывается как アルイワ, а азбучная словарная форма как アルイハ). Азбучно-иероглифическая запись позволяет отличить друг от друга омонимы.

3. *Частеречная информация*. За основу взята система частей речи, принятая в стандартной школьной японской грамматике. Несмотря на отмечавшиеся недостатки школьной системы частей речи и предлагавшиеся варианты ее исправления, авторы корпуса посчитали, что с ней будет проще работать и разметчикам корпуса, и его будущим пользователям. При этом принят гибкий подход к описанию явлений, не укладывающихся в школьную систему. Система частеречной разметки корпуса по необходимости может исправляться и расширяться. Принятая система не настолько дробна, как частеречные системы, используемые в японских автоматических морфологических анализаторах. Однако отказ от более дробного членения частей речи вполне оправдан: дело в том, что за пределами деления на основные части речи у разных исследователей начинаются расхождения в определении частеречных подклассов, и выбрать какую-то одну систему было бы трудно и непрактично.

Определение частеречного класса слова происходит путем исследования контекстов, в которых оно употребляется в корпусе. Для долгих и кратких единиц вместе предусмотрено 15 частей речи: существительные, местоимения, непредикативные прилагательные, неизменяемые прилагательные, наречия, союзы, междометия, глаголы, предикативные прилагательные, служебные глаголы, частицы, приставки (sentooji), суффиксы (setsubiji), символы, запинки (iiyodomi). Приставки и суффиксы, такие как *-gatai*, *-rashii*, выделяются в отдельную часть речи, поскольку обладают в японском большей синтаксичностью, чем суффиксы европейских языков. К символам относятся, например, имена разделов, названные латинскими буквами, или автонимные употребления слов.

4. *Словоизменительный тип* выделяется у предикативных прилагательных (прилагательные на *-i*, *-ku, -shiku* и формы из клас-

сического языка бунго), глаголов, имен *канго,* которые способны сочетаться с глаголом *suru* ‘делать’, и суффиксов, форма которых определяется как адъективная (*-gatai*) или глагольная (*-garu*) в соответствии с грамматическим типом суффикса.

5. *Словоизменительные признаки* присваиваются только словам изменяемых частей речи и определяют форму данной словоформы.

При ручной разметке перечисленные признаки (кроме текстовых границ словоформ) определялись с применением компьютерных программ. В случае слитного произношения слов исходные вероятные формы восстанавливались, и морфологическая разметка давалась уже для них (например, для вспомогательной глагольной формы *-teru* восстанавливаются *-te* и *iru*). Для сокращений наряду с полной формой указывается, что это сокращение. Заполнители относятся к классу междометий.

3.7. Сегментация на синтагмы

Составители корпуса пришли к выводу, что понятие *предложения* плохо применимо к устной спонтанной речи. Ни формальные, ни семантические критерии не позволяют выделить в речи единицы, которые соответствовали бы привычному для письменного текста предложению: выделение предложений по финитным формам глагола или другим признакам конца предложения дает очень длинные единицы, сегментация по паузам дает единицы не всегда имеющие цельную синтаксическую структуру, для выделения семантически целостных единиц трудно подобрать критерии. По мнению авторов, гораздо более осмысленные результаты дает разбиение текста на синтагмы (setsu). Именно такие единицы обладают в устной речи структурной самостоятельностью и достаточной внутренней целостностью и могут быть использованы как минимальные единицы в других видах анализа [Maruyama 2008].

Задача выделения синтагм состоит в нахождении их границ и определении типов самих синтагм. Сегментация текста на синтагмы состояла из автоматического определения границ синтагм и ручной правки полученных результатов. Для первого этапа использовалась программа CBAP (Clause Boundary Annotation Program), которая способна определять границы синтагм на основе грамматической информации (глагольных форм, союзов, локальной морфологиче-

ской информации), а также определять типы выделенных границ. Всего различается 49 типов границ. Они разделены на абсолютные (правая граница синтагмы соответствует концу предложения), сильные (не конец предложения, но разрыв в речи) и слабые (обычно не сопровождаются большими разрывами в речи). Полученные в результате синтагмы делятся на несколько типов по степени их синтаксической и семантической самостоятельности. Эти сведения позволяют предсказывать синтаксическое поведение единиц (сферу действия модальных показателей, свойства тематических и падежных показателей). Особенно интересны с точки зрения лингвистического анализа разбиения по абсолютным и сильным границам.

Приписанный синтагмам тип имеет либо морфологический (синтагмы на *-tari,* на *-tewa,* на *-temo*), либо частеречный (синтагмы глагольного или именного типа), либо лексико-семантический характер (синтагмы причины на *-kara,* причины на *-node,* цитации на *-toiu,* сочинения на *-de,* и др.).

Ручная пост-обработка состояла в том, чтобы исправить те места, где проявления особенностей устной спонтанной речи особенно сильны и не позволили получить надежный результат автоматически. Вот некоторые случаи, потребовавшие ручного исправления:

1. Единица *de* интерпретирована как послелог там, где это связка.

2. Вставка одних синтагм внутрь других. При этом вставленные синтагмы могут иметь внутри себя собственную сильную границу.

3. Спонтанное изменение плана речи, обрыв фразы на середине.

4. Ошибочно интерпретированы как конец синтагмы вставленные внутрь синтагмы заполнители пауз (*nante iu n desu ka* ‘как бы это сказать’), маркеры оговорок (*to iu desu ka* (‘не, не так’), которые зачеркивают сказанное, междометие *ne* ‘не так ли’ и другие единицы.

5. Исправление говорящим сказуемого приводит к появлению в тексте двух глаголов в финитной форме — ошибочного и правильного. Программа ошибочно проводит между ними границу.

6. Проблемы в структуре зависимостей. Тематическая группа (на *wa* или *mo*) может относиться к нескольким синтагмам, разделенным сильной границей, и требуется присоединить отделившиеся тематические группы.

7. Инверсия порядка следования подлежащего и сказуемого.

8. Отсутствие у единицы синтаксического хозяина.

9. Проблемы дискурсивного характера: вставка темы или выражения, подводящего итог сказанному, точка смены темы.

3.8. Дискурсивная разметка

Дискурсивная разметка выполнена для 40 записей из ядра корпуса. Разметка проводилась в терминах теории Б. Грош и К. Сиднер [Grosz and Sidner 1986].

Считается, что дискурсивная цель говорящего получает выражение в поверхностной структуре текста. Определение цели говорящего позволяет понять, почему для ее достижения он выбрал данное речевое поведение и данный способ изложения. В принятом подходе дискурс разбивается на сегменты. Подразумевается, что это разбиение возможно провести без остатка. Задача дискурсивной разметки — определить, какой вклад вносит каждый сегмент в достижение общей цели дискурса. В ходе анализа выделяются сегменты, им даются заголовки (дескрипторы). Минимальной единицей анализа считается бунсэцу (см. выше). Анализ проходит в два этапа: 1) разметка каждого текста тремя разметчиками и 2) обобщение полученной разметки экспертами.

Перед разметчиками ставится задача выделить некоторые целостности, которые можно объединить под одним заглавием и которые осознаются как отдельная тема (wadai). Для каждого сегмента выделяются его начало, конец, определяются его цель, возможно, подцели, добавляются комментарии. В ходе первого этапа разметчик сначала слушает текст один раз и разбивает его на 1-15 частей, указывая неформальным языком их цели. Затем можно слушать текст сколько угодно раз и уточнять полученную разметку. Описание цели намеренно не формализовано, поскольку авторы стремились получить индивидуальные описания и выявить различные взгляды на один и тот же текст: если ввести ограничения, то индивидуальность описания пропадет. Степень дробности разбиения на сегменты не ограничивалась.

Если между разными разметчиками обнаруживается единство в определении границ сегментов, то для полученных сегментов выбирается заголовок, состоящий из двух частей: темы (то, что

объясняет говорящий), и оценочного дескриптора (какими средствами говорящий достигает своей гипотетической цели). Из этих пар затем строится конечный заголовок сегмента, который может содержать служебные слова; при этом допускается перифразирование. Например, рассказ о тонких стенах в общежитии, которые не нравятся говорящему, получает заглавие «проблема тонких стен». Возможные дескрипторы разделены на несколько классов. Главное противопоставление классов связано со степенью субъективности оценки. Среди субъективных выделяются: польза (выгода, недостаток, проблемное место), отношение (приятно, вызывает радость, неприятно), особенность, интерпретация (впечатление, мысль по поводу). Среди менее субъективных: содержание, состояние, вид, сорт, форма, атрибуция, результат. Для описания лекционных записей составлен свой список дескрипторов: определение, состав, объект, принципы, пример, метод, процедура, способ, направление, распределение (в речи о результатах экспериментов).

Если разметчики не сходятся в оценке, то сначала проверяют, насколько удачно определена тема, либо пытаются ввести новый дескриптор с указанием его связей с имеющимися.

Установление целей сегментов состоит из двух этапов:

1. Объединяются результаты работы 3 разметчиков. При этом границы сегментов могут не совпадать. Там, где есть совпадение у двух человек, постулируется граница. Для спорных случаев предусмотрены формальные критерии выбора.

2. Определяется цель дискурса на основе его подцелей. Если подцель всего одна, то она совпадает с целью. Если больше, то делается попытка сначала объединить темы и оценочные дескрипторы подцелей. Часто они либо имеют однотипные части, либо просто совпадают. Если эта процедура не дает результата, то проверяется, не задал ли сам говорящий в начале выступления план или предполагаемое содержание речи. В итоге проводится общая проверка согласованности общих и частных результатов.

Авторы намеренно не использовали в инструкциях конкретные языковые примеры или сведения о паузах, а дали разметчикам свободу выбора, поскольку считают, что необходимо учитывать результаты решений разметчиков, которые те делают

на основе языковой и внеязыковой информации и содержания текста.

3.9. Синтаксическая разметка

Для 500 тыс. слов выполнена синтаксическая разметка в терминах традиционной для японской грамматики системы зависимостей *kakari-uke*, в которой строятся синтаксические деревья непосредственного подчинения. Направлены зависимости от подчиненного к хозяину. Поскольку японский — язык левостороннего ветвления, в большинстве случаев зависимости оказываются направлены слева направо, однако есть исключения. За единицы синтаксического анализа приняты *бунсэцу*.

При синтаксической разметке выявлен ряд проблем, вызванных спецификой устной речи, вот некоторые из них:

1. В случае исправлений ошибочно произнесенная единица не встраивается в общее дерево, а выпадает из него. В таких случаях первоначально произнесенная часть ставится в подчинение исправленной.

2. Вставленные синтагмы имеют свою собственную структуру, не связанную со структурой объемлющей синтагмы, для них структура строится отдельно.

3. Инверсия, когда зависимый член оказывается справа от хозяина, помечается особой связью, идущей справа налево.

4. Отсутствие у единицы синтаксического хозяина помечается особым образом.

В синтаксической структуре выделяются как традиционные типы связей (сочинительная, аппозитивная («президент Кеннеди»), уточняющая («такие [вещи], как мандарины или яблоки»), так и специфические для данного проекта – исправляющая и инвертированная.

Бунсэцу как единицам синтаксической структуры приписываются пометы, например: заполнитель, союз, междометие, обращение, нет хозяина, пересечение (непроективность), старояпонский язык и др.

3.10. Фонетическая разметка

Выполнена для ядра корпуса [Fujimoto 2008]. Разработанная система записи призвана отражать современное состояние языка и от-

слеживать происходящие в нем фонетические изменения. По строгости она занимает промежуточное положение между подробной и упрощенной фонетической транскрипцией. Единицами транскрибирования являются фонемы (bunsetsu-on). Запись сделана буквами латинского алфавита, противопоставление больших и малых букв значимо. В файле каждая единица записывается в строку с указанием времени конца ее звучания. Единицы могут вкладываться друг в друга. Помимо имени фонем используются теги для:

— закрытого участка во время произнесения взрывной согласной или аффрикаты,
— паузы,
— остаточной гласной форманты после окончания колебания голосовых связок,
— колебания связок после гласной,
— гортанного скрипа (voice fry),
— неопределенного гласного,
— неопределенного согласного,
— шума,
— дыхания,
— начала артикуляции.

В случае, когда не удается установить границы звуковой единицы или последовательности единиц, весь комплекс объединяется в одну формальную единицу, которой приписываются все характеристики, определимые для входящих в ее состав элементов. Отдельно описаны сочетания единиц, для которых такое совмещение наиболее вероятно.

Разметка проводилась в 5 этапов:

1. Автоматическое порождение транскрипции по аудиоданным.
2. Выравнивание разметки при помощи алгоритма, основанного на скрытой марковской модели.
3. Приведение разметки вручную к стандартной системе записи.
4. Проверка и исправление двумя специалистами по фонетике.
5. Разрешение проблем, возникших на этапе 4.

Помимо самой фонетической разметки в результате работы получен перечень наиболее проблемных для такой разметки случаев,

а также таблица мор современного японского языка с их фонетической записью.
Для разметки использовалась бесплатно распространяемая программа WaveSurfer.

3.11. Просодическая разметка

Выполнена для ядра корпуса с целью обеспечить возможность поиска единиц по их интонационным характеристикам [Koiso 2003]. За основу взята просодическая транскрипционная система ToBI (Tones and Break Indices), на основе которой для токийского диалекта разработана система J-ToBI. Подробное описание см. в статьях [Igarashi 2008; Maekawa et al. 2002].

3.12. Характеристика речи

Каждому выступлению в ходе записи один из звукооператоров давал субъективную характеристику стиля и степени спонтанности речи. Эта оценка позволяет в целом различать между собой однотипные записи. Оценка проводилась по 5-балльной шкале по следующим параметрам: спонтанность выступления, доля сложных специальных слов, скорость речи, четкость произношения, присутствие диалектных особенностей в лексике или на других языковых уровнях (степень литературности языка), стиль речи. Помимо цифровых позиций в анкетах были предусмотрены словесные оценки (речь беглая или нет, монотонная, выразительная, расслабленная, напряженная и др.).

4. Распространение корпуса

Корпус распространяется на 18 DVD дисках, куда включены аудиозаписи, транскрипция, разметка всех описанных выше типов, рефераты текстов, словарь всех кратких единиц, встречающихся в корпусе, сведения о говорящих (пол, возраст, место рождения, краткая биография), инструкция пользователя, программные инструменты для работы с корпусом (для поиска записей, для прослушивания записей, их анализа). Текстовые данные переведены в формат XML. Для поиска по ним используются средства XPath, поисковый запрос можно составлять при помощи визуального конструктора, который позволяет задавать критерии поиска

и способ представления найденных данных. Данные, полученные в результате запроса, можно сохранить в файл в формате CSV (таблица, в которой значения ячеек разделены запятыми).

Примеры аудиозаписей и разметки текстов можно найти по адресу *http://www.kokken.go.jp/katsudo/seika/corpus*. Там же регулярно публикуется новая информация по проекту.

Литература

[Fujimoto 2008] Fujimoto Masako. «Nihongo hanashikotoba koopasu» no bunsetsuon joohoo (Сведения о сегментных фонетических единицах в Корпусе японской разгноворной речи). // Nihongo gaku, 2008, Vol.27-5, pp.90–102.

[Gengo 2006] Gengo. Tokushuu: kotoba no tan'i (Журнал «Язык». Тематический выпуск «Языковые единицы»). 2006, Vol.35, No.10.

[Grosz and Sidner 1986] Grosz, B.J., Sidner, C.L., Attention, Intentions, and the Structure of Discourse // Computational Linguistics, 12:3, 1986.

[Igarashi 2008] Igarashi Yosuke. «hanashikotoba koopasu» no inritsu joohoo (Просодическая информация в Корпусе японской разговорной речи) // Nihongo gaku, 2008, Vol.27-5, pp.103–113.

[Koiso 2003] Koiso Hanae. Koopasu ni yoru onsei danwa no kenkyuu (Исследование устных диалогов при помощи корпуса) // Nihongo gaku, 2003, Vol.22, pp.200–209.

[Maekawa et al. 2001] Maekawa Kikuo, Kikuchi Hideaki, Kagomiya Takayuki, Yamaguchi Masaya, Koiso Hanae, Ogura Hideki. «Nihongo hanashikotoba koopasu» no koochiku ni okeru keisanki riyoo (Использование вычислительной техники при создании Корпуса японской разговорной речи) // Nihongo gaku, 2001, Vol.20, pp. 61–79.

[Maekawa et al 2002] Maekawa Kikuo, Kikuchi Hideaki, Igarashi Yosuke, Venditti Jennifer. X-JToBI: an Extended J-ToBI for Spontaneous Speech // ICSLP, 2002, pp.1545-1548.

[Maekawa 2008] Maekawa Kikuo. «Nihongo hanashikotoba koopasu» no sekkei to jisso (Корпус японской разговорной речи: план и его реализация) // Nihongo gaku, 2008, Vol.27-5, pp.54–62.

[Maruyama 2008] Maruyama Takehiko. «Nihongo hanashikotoba koopasu» no setsu tan'i joohoo (сведения о синтагмах в Корпусе японской разговорной речи) // Nihongo gaku, 2008, Vol.27-5, pp.82–89.

[Nihongo 2006] Nihono hanashikotoba koopasu no koochiku hoo. Kokuritsu kokugo kenkyuujo hookoku 124 (Метод построения корпуса разговорной японской речи. Отчет Гос. института национального языка). Токио: 2006. (*http://www.kokken.go.jp/katsudo/seika/corpus/csj_report*)

[Ogura 2008] Ogura Hideki. «Nihongo hanashikotoba koopasu» no gengo tan'i (Языковые единицы в Корпусе японской разговорной речи) // Nihongo gaku, 2008, Vol.27-5, pp.72–81.

[Uchimoto 2003] Kiyotaka Uchimoto, Kazuma Takaoka, Chikashi Nobata, Atsushi Yamada, Satoshi Sekine, Hitoshi Isahara. Morphological Analysis of the Corpus of Spontaneous Japanese. In Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003.