

**Решения и ответы**

**Блок 1. «Теория языка».**

**Вопрос 2.**

В задаче предлагается рассмотреть средства личной референции. Таковых - местоимений с личной референцией - в примере три - кореферентные *towmtu* и *žu* и *was*. Попробуем ответить на следующий вопрос - имеем ли мы дело с прямой или косвенной речью? Использование в цитате местоимение второго лица, казалось бы, указывает на то, что цитирование прямое. Действительно, в закавыченной цитате в переводе употреблено местоимение второго лица, то есть адресат исходного речевого акта кодируется так же, как в исходном речевом акте. (Теоретически можно было бы предположить, что адресат текущего речевого акта и адресат исходного речевого акта совпадают - но это крайне маловероятная ситуация - тогда получилось бы, что речь третьего лица пересказывается тому же человеку, к которому она была обращена.) Если сдвига дейктического центра не происходит, следовательно, это прямая речь.

С другой стороны, в той же клаузе используется логофорическое местоимение. Логофорическое местоимение, кажется, указывает на косвенно-речевой характер цитирования - по определению, логофорическое местоимение используется для выражения кореференции с субъектом матричного глагола речи, и замещает исходное местоимение первого лица.

(Данное явление допускает несколько теоретических интерпретаций. С одной стороны, говорят, что в такой ситуации мы имеем дело с полукосвенной речью - *semidirectness*. Однако для этого конкретного языка, как и для некоторых других языков, правильнее было бы говорить, что косвенная стратегия в нем вообще отсутствует, и логофорическое местоимение играет особую функцию - не выражение кореференции с исходным говорящим, а выражение несовпадения с текущим говорящим).

**Вопрос 3.**

В задаче предлагается найти функциональное, внеязыковое объяснение обсуждаемым фактам. Как кажется, тут есть два возможных решения.

(а) Первое связано с актуализованностью в сознании носителей европейских языков образа циферблата. Традиционный циферблат явным образом выделяет начало часа, половину часа и четверти часа. Несколько менее сильно выделены промежуточные пятиминутки (двенадцатые часа). При этом решении остаются необъясненными восьмые часа.

(б) Второе, в целом почти столь же привлекательное решение, связано с делением пополам. Действительно, получасы, четверти и восьмые получаются путем деления временного отрезка пополам. В данном случае необъясненными остаются двенадцатые доли.

Вероятно, чтобы объяснить все наблюдаемые системы, требуется привлечь оба принципа. Для объяснения двенадцатых долей часа (пятиминуток) вряд ли можно предложить что-либо, кроме циферблата; с другой стороны, для объяснения восьмых часа необходимо привлечь деление пополам (отметим, что оно дает дробное число минут). При этом каждый из редких случаев отмечается лишь в одном языке (каждый в своем).

## Блок 2. «Компьютерная лингвистика».

### Вопрос 2.

Представьте себе, что вы хотите приготовить на ужин что-то особенное. Вы смотрите, что у вас есть в холодильнике, а потом начинаете поиск в интернете. Там находится такое огромное количество разных рецептов для ваших продуктов, что если читать и обдумывать их всех – вы точно останетесь голодным. И потом – вот отличный рецепт для говядины, которой у вас нет, а можно ли ее заменить свининой, которая у вас есть?

Придумайте и опишите, как бы мог быть устроен сервис, который бы помогал пользователю находить те рецепты, которые бы ему нужны, упорядочивал бы близкие рецепты, и давал представление о том, какие продукты необязательны, а какие можно заменить. В вашем ответе должны быть четко указаны, какие именно данные вы будете использовать, как вы их будете собирать, и обрабатывать; какие лингвистические проблемы у вас могут возникнуть и как вы их будете решать. Наконец, вы должны описывать, как вы будете проверять, насколько хорошо работает ваша система.

В сервисе должны быть реализованы три функциональные задачи

- 1) Поиск релевантных рецептов
- 2) Объединение похожих рецептов в тематические группы
- 3) Выделение обязательных/необязательных компонентов

Возможное решение

- 1) Сбор корпуса рецептов
- 2) Создание грамматики рецептов. Выделение ключевых элементов:
  - а) название основного блюда или промежуточных состояний (бульон, смесь, отвар)
  - б) ингредиенты (продукты, специи, вода и т.д)
  - в) действия (налить, перемешать, запекать и т.д.)
  - г) объем (литры, столовые ложки, и т.д.)
  - д) время действия( 15 минут, до готовности)и отношений между ними:  
например,  
действие (ингредиент 1 (объем), ингредиент2 (объем), время) > продукт1  
действие (продукт1, время) > продукт2
- 3) Разметка корпуса, аннотирование элементов рецепта и отношений. Разметка может осуществляться с помощью словарей и эвристических правил, и с помощью машинного обучения
- 4) Лемматизация корпуса
- 5) Извлечение структурированной информации из каждого рецепта, соотнесение структурированного фрейма с текстом документа. Нормализация объема всех ингредиентов на 2 порции.
- 6) Кластеризация семантических структур, выделение похожих рецептов и параметров их варьирования. Порог варьирования определяет близость

рецептов и возможность замены компонентов или необязательных ингредиентов

- 7) Поиск по кластерам семантических структур рецептов, разработка формата поисковой выдачи (например, инфографика)

#### Лингвистические проблемы

- 1) Связанные с нормализацией – решение с помощью словарей и регулярных выражений:
  - a. Нормализация уменьшительных названий ингредиентов и блюд (картошка/картофель, оладьи/оладушки)
  - b. Нормализация мер объема и времени (мин/минуты, ст.л./столовая ложка)
  - c. Нормализация глаголов и их дериватов (мешать/перемешать/смешать)
- 2) Связанные с лексическими отношениями – решение с помощью нормализации повторяющихся вариативных компонентов близких рецептов
  - a. Синонимией: хлеб/булка, сахар/песок
  - b. Гиперонимией мясо/говядина
  - c. Омонимией салат(овощ)/салат (блюдо)
- 3) Связанные с кореферентностью или бриджингом - решение с помощью эвристических правил или машинного обучения:
  - a. Употребление анафорических местоимений (замешайте тесто и положите его в холодильник)
  - b. Употребление бриджинга: (Достаньте говядину. Нашпигуйте мясо чесноком.)

Проверка работы системы должна учитывать параметры точности и полноты. Ответы системы должны сравниваться с эталонными ответами предоставляемыми экспертами, кроме точности выдачи, должна учитываться полнота выдачи – все ли возможные ответы предоставлены системой.

Предложенное решение является лишь одним из возможных подходов. В оценке задачи учитывалось, во-первых то, насколько продуманной была работа с корпусом рецептов и со структурными элементами данных, насколько отрефлексированы были лингвистические проблемы, все ли параметры проверки системы были учтены.