

ОБРАБОТКА ОМОГРАФОВ В СИСТЕМАХ АВТОМАТИЧЕСКОГО СИНТЕЗА РЕЧИ НА АНГЛИЙСКОМ ЯЗЫКЕ

Ю. Н. Цыганкова

Белорусский государственный университет, г. Минск;

ulistia@yandex.by;

науч. рук. – В. В. Воронович

В статье описываются возможные алгоритмы обработки входной информации автоматическими синтезаторами речи. Особое внимание уделяется проблеме озвучивания автоматическими синтезаторами омографов английского языка. На основании проведенного исследования выявляется правильность озвучивания омографов, а также описываются проблемы, возникающие в процессе озвучивания.

Ключевые слова: омонимия; омограф; автоматический синтез речи; система автоматического синтеза; синтез текст-в-речь.

Известно, что английский язык богат на наличие омографов. Увеличение их количества в современном английском языке обусловлено историческим развитием языка – это один из факторов, отличающих современный английский язык от его древнего варианта. Однако проблема омографии в английском языке изучена недостаточно на текущий момент.

Омография как явление вытекает из более обширного явления – омонимии. Среди омонимов выделяются: омографы (лексические единицы, одинаковые графически), омофоны (лексические единицы, одинаковые фонетически) и омографы-омофоны (лексические единицы, которые совпадают в своем как графическом, так и фонетическом оформлении). Например, *flour* ‘мука’, *flower* ‘цветок’; *peace* ‘мир’, *piece* ‘кусок, часть’; *sell* ‘продавать’, *cell* ‘клетка’ – омофоны; *tear* ‘рвать, резать’, *tear* ‘слеза’; *bow* ‘лук, бант’, *bow* ‘поклон’ – омографы; *bear* ‘1. медведь, 2. терпеть, 3. Носить’; *fair* ‘1. честный, 2. красивый, 3. Ярмарка’; *row* ‘1. ряд, 2. грести’ – омографы-омофоны.

В русском языке одной из причин возникновения омографов и, пожалуй, основной, является различное ударение: [з^амок] – [зам^ок] – самый очевидный и часто встречающийся пример. Проблема омографии возникает и в случае, когда в тексте не проставлены точки над “ё”, например: *небо* [н^ебо] и [н^ёбо], *осел* [ос^ел] и [ос^ёл], *сел* [сел] и [сёл], *мел* [мел] и [мёл]. «В русском языке существуют два источника графической омонимии: вариативность словесного ударения, местоположение которого в письменной речи не указывается (СУ- омографы), и письменная традиция необязательного проставления необходимых точек на букве «Ё» («Ё»-омографы)» [1]. Омографы других типов встречаются редко в

русском языке. В качестве примера омографии, обусловленной иными явлениями, можно привести возникновение омографов из-за различного прочтения буквы “ч”: *У него, сердечного, сердечная болезнь* или *Конечно, это множество конечно* (оба раза в первом случае звук [ш], во втором – [ч’]).

В иностранных языках рассмотрение омографии, как правило, не связано с изменением ударного слога в словах, западное языкознание рассматривает слова одинакового написания с одинаковым ударным слогом, но различным произношением. «Кроме того, явление омографии рассматривают наравне с омонимией, когда произношение слов одинаковое, но смысл разный» [2]. Так как в нашей работе мы рассматриваем проблему озвучивания омографов английского языка автоматическими синтезаторами речи и ошибки, возникающие по причине их неправильного озвучивания, то омографы, вариация произношений которых основана на принадлежности этих омографов к различным частям речи, интересуют нас в первую очередь. В ряде случаев формы существительного и глагола при одинаковом написании различаются произношением отдельных звуков, а не только местом ударения: *insult* ‘оскорбление’ – *insúlt* ‘оскорблять’, *éxport* ‘экспорт’ – *expórt* ‘экспортировать’, *désert* ‘пустыня’ – *desért* ‘покидать’, *use* [ju:s] ‘польза’ – *use* [ju:z] ‘использовать’.

Очень часто в английском языке такие слова могут являться более чем одной частью речи (имя существительное, глагол, имя прилагательное и т. д.). Стоит обратить внимание на то, что почти в каждом таком случае существительное ударяется на первом слоге, а глагол – на втором. Как правило, изучая английский язык, мы узнаем об этой закономерности еще в начале своего пути обучения, поскольку это важный аспект, имеющий влияние на понимание высказывания.

Существуют случаи, когда омографы не меняют свое ударение, а отличаются лишь качеством произношения того или иного звука. Например, *abuse* ‘оскорблять (гл.); злоупотребление (сущ.)’ и *use* ‘использовать (гл.); склонность (сущ.)’. Пары этих омографов различаются произношением ‘s’: в форме глагола он произносится как [z], а в форме существительного он оглушается и звучит как [s]. Также одной из частей речи в паре омографа-глагола может быть прилагательное.

Для тестирования систем автоматического синтеза мы разработали словарь, в котором собрано более сотни омографов английского языка, а также примеры их употребления.

Синтез речи – в широком смысле – восстановление формы речевого сигнала по его параметрам; в узком смысле – формирование речевого сигнала по печатному тексту. Основными средствами воспроизведения

текстовой информации в системах автоматического синтеза являются системы CDS (Common Data Speaker), TTS (Text-to-speech). Среди существующих методов синтеза речи по тексту выделяют параметрический синтез, компиляционный и синтез речи по правилам [3]. На сегодняшний день в область автоматического синтеза речи активно внедряются нейронные сети.

В нашей работе мы сделали выборку из наиболее частотных омографов английского языка. В наш словарь вошло сто омографов. Далее мы просинтезировали эти омографы с помощью выбранных 5 синтезаторов речи для выяснения того, насколько же правильно автоматический синтезатор может справиться с распознаванием и озвучиванием омографа. Примеры предложений, которые мы прорабатывали:

- *I believe I got the **content** ['kɒntent] of his speech. – Я думаю, я понял суть его речи.*

I am **content** [kən'tent] with his speech. – Я доволен его речью.

- *The bullets are made of **lead** [led]. – Пули сделаны из свинца.*

A good manager has to **lead** [li:d] a team. – Хороший менеджер должен вести команду.

Дополнительно хотелось бы отметить, что из использованных нами ресурсов далеко не все достойно и качественно озвучивают текст даже просодически, что является также серьезным вопросом, так как очень сложно воспринимать на слух предложение, которое неверно проинтонировано, а еще сложнее – предложение, в котором нет завершенности (точка в конце предложения никак не принимается во внимание одной из систем синтеза). На рисунке представлены системы автоматического синтеза речи, которые были отобраны для эксперимента, а также правильность озвучивания ими предложений с омографами.

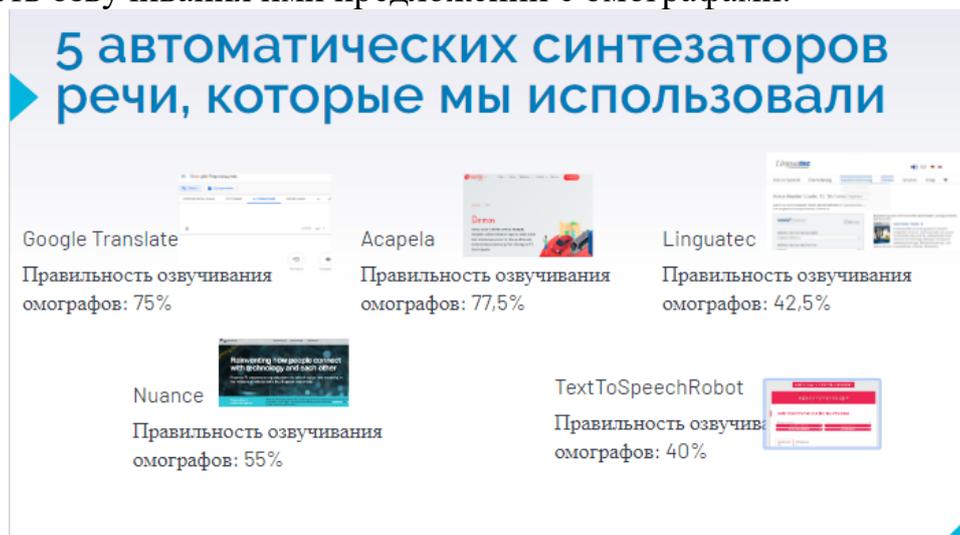


Рис. Правильность озвучивания омографов автоматическими синтезаторами речи

Использованные для тестирования синтезаторы речи направлены на коммерческое озвучивание: использование синтезаторов речи в сфере обслуживания (общение с клиентами), автоответчики, озвучивание текстовой информации на экранах электронных устройств для людей с дефектами зрения, также для людей с дислексией, и т.д. Проблема омографии подобными синтезаторами речи не решается в достаточной степени. Не хватает ресурса, который был бы направлен на решение подобной лингвистической задачи.

Библиографические ссылки

1. *Лобанов Б. М.* Проблема разрешения «ё»-омографов при синтезе речи по тексту [Электронный ресурс]. – Режим доступа: <http://www.dialog-21.ru/digests/dialog2009/materials/html/45.htm> – Дата доступа : 22.03.2020.
2. *Чемерилов В.В., Фадеев А.С.* Система автоматического разрешения омографии на основе семантической связи слов смежных предложений в текстовом отрывке [Электронный ресурс]. – Режим доступа: <https://cyberleninka.ru/article/n/sistema-avtomaticheskogo-razresheniya-omografii-na-osnove-semanticheskoy-svyazi-slov-smezhnyh-predlozheniy-v-tekstovom-otryvke/viewer> – Дата доступа : 02.02.2020.
3. *Лобанов, Б.М., Цирульник Л.И.* Компьютерный синтез и клонирование речи. Минск : Белорусская наука, 2008.
4. *Смирницкий А.И.* Лексикология английского языка. Москва: Издательство литературы на иностранном языке, 1956 г.
5. *Cambridge dictionary online* [Электронный ресурс]. – Режим доступа: <https://dictionary.cambridge.org/dictionary/> . – Дата доступа : 15.03.2020.
6. *Oxford Learner's Dictionary* [Электронный ресурс]. – Режим доступа: <https://www.oxfordlearnersdictionaries.com>. – Дата доступа : 04.03.2020.