Анализ текстов SMS-сообщений с целью повышения качества их автоматического озвучивания

An analysis of SMS texts aimed at inproving their automatic reading

Людовик Т. В. (tetyana.lyudovyk@gmail.com)

Международный научно-учебный центр информационных технологий и систем НАН Украины и МОН Украины, Киев, Украина

Статья посвящена выявлению особенностей SMS-текстов, препятствующих качественному автоматическому озвучиванию SMS-сообщений. Основные проблемы связаны с неправильным распознаванием языка SMS-сообщения, использованием при написании SMS суржика и сленгов, нестандартной транслитерации и орфографии.

1. Введение

Решение актуальных задач автоматического распознавания и синтеза речи связано с учетом особенностей языковой ситуации, сложившейся в сферах применения этих речевых технологий. Задачи усложняются, если приложения связаны с живым разговорным языком.

Большинство разработанных систем синтеза речи [1–4] озвучивают правильные с точки зрения орфографии и синтаксиса тексты на литературных языках. В условиях многоязычности возникает проблема распознавания языка, на котором написаны тексты.

В данной статье анализируется опыт использования системы синтеза украинской речи для озвучивания SMS-сообщений, отправляемых с мобильных телефонов на стационарные [5]. Большинство текстов SMS-сообщений являются спонтанными, и их автоматическое озвучивание требует учета не только лингвистических, но и социальных, демографических и психологических факторов.

Помимо спонтанности тексты SMS-сообщений отличаются несоблюдением норм литературного языки, использованием сленгов с нефиксированным составом словарей и экспрессивностью. Задачу озвучивания SMS-сообщений в Украине осложняет также сложившаяся языковая ситуация, характеризующаяся двуязычием и наличием смешанного украинско-русского языка — суржика.

По данным переписи населения Украины 2001 года этнический состав страны, определявшийся моноэтнической самоидентификацией взрослых и собственной идентификацией ими своих детей, насчитывал 77,8 % украинцев, 17,3 % русских и 4,9 % представителей других этнических групп. При этом украинский язык назвали родным 67,5 % населения Украины, а русский — 29,6 % украинских граждан (по данным Госкомстата Украины [6]).

В рамках всеукраинских социологических опросов, проведенных Киевским международным институтом социологии в 1991–2003 годах (около 173 000 интервью) [7] проводились устные интервью, целью которых было определить, на каких языках общается взрослое население (старше 18 лет). На вопрос о том, какой язык является наиболее удобным при общении, 47 % граждан Украины назвали украинский и 53 % — русский.

Однако, самоидентификация языка респондентами не отражает их фактического речевого поведения. Проведенные Киевским международным институтом социологии опросы ценны лингвистическими экспертными оценками, данными по окончании интервью интервьюерами-билингвами. Эти оценки свидетельствуют о том, что довольно значительная часть респондентов разговаривала с интервьюерами на смеси украинского и русского языков. Так, в 2000–2003 годах 38,7 % взрослого населения Украины говорило на украинском языке, 46,6 % — на русском и 14,7 % — на суржике.

Часто во время проведения социологических опросов суржикоязычное население фиксируется как украиноязычное. Данные приведенных опросов косвенно это подтверждают: использование суржика зафиксировано у 14 % этнических украинцев и у 5 % этнических русских. Таким образом, озвучивание SMS-сообщений, написанных на суржике, должно осуществляться системой синтеза украинской речи.

2. Цель работы

Целью работы является исследование и учет особенностей SMS-сообщений для более качественного их озвучивания системой синтеза речи. Необходимо классифицировать выявленные проблемы, определить наиболее важные из них с точки зрения влияния на качество синтезированной речи, и выяснить, возможно ли их решение на данном этапе. Для проблем, поддающихся решению, необходимо предложить пути решения.

3. Исследованный материал

Материалом для проведенных исследований послужили реальные SMS-сообщения, отправленные с мобильных телефонов в периоды с 17 мая по 6 июня и с 19 по 27 июля 2008 года. Эти SMS-сообщения были анонимизированы путем исключения данных об отправителях и получателях. Таким образом, анализировались лишь тексты сообщений. Экспертами было выделено пять категорий SMS-текстов в соответствии с языком, на котором они написаны. В таблице 1 представлено распределение проанализированных SMS-сообщений. К категории «условно-украинский язык» отнесены:

- SMS, язык которых невозможно однозначно определить, поскольку все слова текста принадлежат как русскому, так и украинскому языкам (хотя, возможно, произносятся по-разному (ср. «напиши» ([нап'ишы] и [напышы]), «день ([д'эн'] и [дэн']));
- SMS, написанные на суржике.

В данной работе использовались также результаты тестирования компонента автоматического распознавания языков сервиса SMS2Voice [5]. В рамках этого сервиса, работающего в Украине, все SMS-сообщения, отправляемые с мобильных телефонов на стационарные, на первом этапе обработки автоматически классифицируются по языку. В список возможных языков включены только русский, украинский и английский; в случае, если язык определить не удается, по умолчанию языком SMS-сообщения считается русский. После распознавания языка сообщения оно озвучивается соответствующей моноязычной системой синтеза речи.

Автоматическое распознавание языка 5480 SMS-сообщений дало следующие результаты: на русском языке написано 73 % SMS-сообщений, на украинском — 14 %, на английском — 4 %. В 10 % случаев язык определить не удалось.

Таблица 1. Распределение SMS-сообщений по языковым группам

Язык	Количество SMS (17.05.2008–6.06.2008)	Количество SMS (19.07.2008–27.07.2008)	Примеры
Русский	7640 (70,6 %)	3746 (68,4 %)	*Я на работе.*
Украинский	1523 (14,1 %)	843 (15,4 %)	*Як справи? (Как дела?)*
Условно-украинский	1262 (11,7 %)	586 (10,7 %)	*Напиши SMS.* *Умееш находить похидну (производную) с корня?*
Другие языки и нетекстовые SMS	400 (3,7 %)	305 (5,6 %)	*Tutto ok. E mille grazie @1@2@3@*
Всего	10 825 (100 %)	5480 (100 %)	

Таблица 2. Сравнение результатов распознавания языка SMS-сообщений

Язык	Автоматическое распознавание	Экспертное распознавание
Русский	3998	3516
Украинский	748	665
Английский	200	138
Язык не определен	534	
Условно-украинский		945
Другие языки и нетекстовые SMS		216
Всего	5480	5480

4. Многоязычность текстов SMS-сообщений

Проведенное сравнение результатов автоматического распознавания языка SMS-сообщений с результатами распознавания языка этих же SMS-сообщений экспертами-лингвистами дало результаты, приведенные в таблице 2.

Эксперты отнесли к русскому языку существенно меньшее количество SMS-сообщений. Это может быть объяснено тем, что автоматическое распознавание связано с поиском слов в словарях, и если слова текста SMS обнаруживаются как в русском, так и в украинском словарях, то автоматически принимается решение о русском языке. Однако сочетание обнаруженных слов, учитываемое экспертами, может однозначно свидетельствовать в пользу украинского языка. Например: «*я тебе люблю!*». В дальнейшем при распознавании языка SMS-сообщения необходимо учитывать не только наличие слов в словарях, но и их частотность в каждом из языков, ср.: «*як там?*» («*как там?*»).

Существенно меньшее количество SMS, распознанных экспертами как англоязычные, связано с тем, что не учитываются другие неславянские языки. Среди других языков экспертами были выявлены следующие: французский, немецкий, испанский, итальянский, турецкий, вьетнамский, нидерландский, молдавский/румынский, венгерский и грузинский. В настоящее время вопрос озвучивания SMS на этих языках не решен.

В целом, нетранслитерированные орфографически правильно написанные SMS-сообщения на украинском, русском и английском языках не представляют значительных трудностей при их озвучивании.

5. Особенности SMS-сообщений, написанных на украинском и условно-украинском языках

В дальнейшем проводился анализ текстов SMS, отнесенных к украинскому и условно-украинскому языкам. Из категории «условно-украинский язык» были исключены тексты, которые могут быть отнесены к русскому языку, например, «*6 вагон.*», «*Я буду завтра.*». Поскольку SMS на суржике должны озвучиваться системой синтеза украинской речи, украинские и условно-украинские SMS были объединены в одну категорию «SMS на украинском языке». Всего в дальнейшем анализировалось 1529 SMS-сообщений.

Наличие словаря литературного языка, словаря ненормативной лексики, словаря молодежного и SMS-сленгов, а также таблиц транслитерации позволяет использовать для разграничения SMS-текстов критерии, перечисленные в таблице 3. Наиболее сложно разграничить SMS, написанные на литературном украинском языке с орфографическими ошибками (описками), и SMS, написанные на суржике.

К группе «SMS на литературном языке с нестандартной орфографией» отнесены SMS с сокращениями слов.

В таблице 4 приведено распределение SMS-сообщений по группам.

Как правило, озвучивание орфографически правильно написанных на литературном языке SMS-сообщений не представляет затруднений.

Таблица 3. Критерии разграничения SMS-текстов

Критерии разграничения	Примеры
Все слова текста SMS-собщения найдены в словаре литературного языка непосредственно или после обратной транслитерации	*Вітаю з днем народження!* *Ja vzhe na misci.*
Прочитанное вслух SMS-собщение с нестандартной орфографией или нестандартной транслитерацией звучит на литературном языке	*ЯНА ЗВАРЫ БОРЩ И ВИДРО ВАРЕНЫКИВ*
Хотя бы одно слово текста SMS входит в список ненормативной лексики	*Бачу тобі …!*
Хотя бы одно слово текста SMS написано на молодежном или SMS-сленге	*Я в унівєрі. Пліз.:-(*
Все остальные SMS считаются написанными на суржике	*Визивай пожарних на склади загорилися ящики.*

Таблица 4. Распределение SMS-сообщений, написанных на украинском и условно-украинском языках

Категории SMS	Количество SMS
SMS, орфографически правильно написанные на литературном языке	663
SMS на литературном языке с нестандартной орфографией и/или транслитерацией	590
SMS с использованием ненормативной лексики	42
SMS с использованием молодежного и/или SMS-сленга	20
SMS на суржике	214
Всего	1529

6. SMS-сообщения с нестандартной/неоднозначной транслитерацией

Значительная часть исследованных SMS-сообщений (17 %) написана с использованием латинского алфавита, что обусловлено экономией времени и денег, а также возможностями кодировок. В более половины случаев игнорируется наличие официальных таблиц транслитерации.

Наиболее часто отклонения происходят при транслитерации йотированных гласных букв и сочетаний «йотированная гласная + "й"» («*lublu*» вместо «*lyublyu*», «*daite*» вместо «*dajte*»). По-разному транслитерируются «ж», «ш», «ц», «щ», «с». Часто при транслитерации исчезают мягкий знак и апостроф. Широко используются цифры 1 и 4, реже — цифры 0, 3, 6.

Справедливости ради следует отметить, что таблицы транслитерации неудобны, и даже мобильные операторы, посылая SMS-сообщения, не соблюдают правила транслитерации.

Часто одна и та же латинская буква соответствует различным украинским; иногда это случается в одном и том же слове, например, «*rozbudulu*» (латинская «u» используется как в качестве «и», так и в качестве «у»).

В настоящее время учет нестандартной транслитерации латиницей осуществляется следующим образом. Для каждого транслитерированного слова порождается множество вариантов записи кириллицей с учетом вероятности замен латинских букв кириллическими. Затем все варианты в порядке убывания вероятности проверяются на наличие в словаре.

Особую группу составляют украинские SMS-сообщения, транслитерированные русскими буквами, например, «*Який ти писля цёго друг.*», «*Я для неї вірш напысав.*», «*А ты де на Днипри?*». Подобные SMS либо распознаются как русскоязычные и озвучиваются системой синтеза русской речи, либо распознаются как украиноязычные и озвучиваются по-украински с ошибками.

7. SMS с нестандартной орфографией

Нарушение норм пунктуации при написании SMS-сообщения в данной работе не рассматривается, поскольку оно в значительно меньшей степени влияет на качество озвучивания, чем нарушение норм орфографии. Нестандартная орфография зафиксирована в 39 % украинских SMS (не считая SMS, написанных на суржике), что представляет собой одну из главных проблем при озвучивании.

Отклонения от стандартной орфографии могут быть намеренными («*Яаа лююблю тильки тебеее!*»), в результате описок («*на дороагах*») и неграмотности («*Візьми тіліфон.*»).

Несмотря на то, что при опросах 73 % респондентов-украинцев заявили, что хорошо владеют письменным стилем украинского языка, и 71 % — устно-разговорным, самооценка не соответствует реальному уровню владения языком. Не может не тревожить неграмотность молодежи.

Некоторые часто встречающиеся в текстах SMS орфографически неверно написанные слова и словосочетания внесены в словарь и в настоящее время озвучиваются правильно, например, «*будь ласка*», «*будь-ласка*», «*будьласка*». Необходим более планомерный подход к проверке орфографии с учетом того, что часто соседние слова пишутся слитно, а одно слово может быть разделено на части.

8. SMS с нестандартными сокращениями

Нестандартные сокращения в SMS-текстах не могут быть расшифрованы. Как правило, по тексту, состоящему из одного-двух предложений, сложно восстановить семантический контекст. Неясно, когда системы синтеза речи будут в состоянии определить, что в тексте «*Вона вже говорила, шо ти весь час пропускаеш к. р.*» речь идет о школьных контрольных работах, а в тексте «*З ДН тебе*» — о дне рождения.

9. SMS с использованием элементов сленгов

Молодежный и SMS сленги отличаются динамичностью, постоянным появлением новых слов и выражений. Очевидно, для озвучивания соответствующих SMS-сообщений необходимы специальные словари.

Другой отличительной особенностью этой группы SMS-сообщений является экспрессивный характер. Озвучивание SMS-приколов, возможно, требует особой выразительной просодики. Это должно быть учтено на этапе разработки речевой базы данных при составлении текстов, на основе которых производятся акустические записи голоса диктора. Диктор, чей голос впоследствии будет использован для озвучивания SMS, должен читать эти тексты с подчеркнутой выразительностью.

10. SMS с использованием ненормативной лексики

Существует список слов ненормативной лексики. При озвучивании слова из этого списка заменяются сигналом «би-и-п». К сожалению, зафиксировать список не удается. Некоторые пользователи упражняются до тех пор, пока, видоизменяя неприличные слова, не добьются желаемого звучания.

11. SMS на суржике

Как показывают результаты социолингвистического мониторинга, на суржике общаются 15 % взрослого населения Украины и 27 % студентов.

В данной работе принято рабочее определение суржика как смеси украинского и русского языков. Для удобства анализа разделение украиноязычных SMS-сообщений произведено на непересекающиеся группы. Группа «SMS на суржике» содержит только те сообщения, которые не вошли в остальные группы (см. таблицы 3 и 4). В действительности, элементы суржика могут присутствовать в одном SMS-сообщении наряду с ненормативной лексикой и молодежным сленгом. По нашим данным, на суржике пишется от 14 % до 18 % украиноязычных SMS-сообщений.

Как правило, лексика в суржике взята из русского языка, а большая часть грамматики — из украинского. Простые случаи, когда русские слова встречаются в текстах на суржике в неизменном виде («*срочно*», «*тоже*»), могут быть учтены с помощью расширения украинского словаря. Дополнительное неправильное орфографическое написание («*нада*», «*пожалуста*») усложняет задачу.

Более сложные случаи взаимопроникновения языков («*відказуємося*», «*задержуюсь*») нуждаются в дополнительном анализе. Существует мнение, что суржик — это индивидуальное нарушение языковых норм, что он всегда персонален и, следовательно, не может быть закреплен в словарях и грамматиках. Однако, в соответствии с другой точкой зрения, суржик возникает в результате системной интерференции. В настоящее время суржик в основном рассматривается как испорченный русизмами украинский язык. Возможно, в будущем он получит иной статус и будет составлен словарь суржика. Аналогично, будут разработаны системы синтеза речи на суржике.

12. Выводы

Языковая ситуация в Украине изучена недостаточно. Фиксация суржикоязычного населения как украиноязычного не отражает реального соотношения между языками общения.

Учет прагматических факторов, влияющих на написание SMS сообщений, способствует более точному автоматическому распознаванию языка сообщения.

Основные проблемы, возникающие при озвучивании текстов SMS, связаны с нестандартными транслитерацией и орфографией, использованием суржика и сленгов.

Для озвучивания SMS, написанных на распространяющемся быстрыми темпами суржике, необходимо либо расширение речевых баз данных общеукраинского языка, либо создание отдельных, специальных речевых баз данных суржика. Последнее не исключено, если суржик будет признан самостоятельным языком общения.

Литература

1. *Quazza S., Donetti L., Moisa L., Salza P. L.* ACTOR: a Multilingual Unit-Selection Speech Synthesis System // 4th ISCA Tutorial and Research Workshop on Speech Synthesis. 2001. Paper 209.
2. *Лобанов Б. М., Цирульник Л. И.* Компьютерный синтез и клонирование речи // Минск: Белорус. наука, 2008. 337 с.
3. *Oparin I., Talanov A.* Outline of a New Hybrid Russian TTS System // Proc. of the 12th International conference on Speech and Computer SPECOM 2007. Moscow.: 2007. Pp. 603–608.
4. *Romsdorfer H., Pfister B.* Text analysis and language identifcation for polyglot text-to-speech synthesis // Speech Communication, 2007. Vol. 49, pp. 697–724.
5. *Lyudovyk T., Brozinski S., Noner M., Robeiko V., Sazhok M.* Speech Synthesis Applied to SMS reading // Proc. of the 13th International Conference «Speech and Computer: SPECOM'2009». — St. Petersburg: 2009. Pp. 300–305.
6. http://www.ukrcensus.gov.ua/results/general/language/
7. *Хмелько В. Є.* Лінгво-етнічна структура України: регіональні особливості і тенденції змін за роки незалежності // Наукові записки НаУКМА. «Соціологічні науки». К.: 2004. Т. 32. С. 3–15.