АНАЛИЗ ТЕКСТОВ SMS-СООБЩЕНИЙ С ЦЕЛЬЮ ПОВЫШЕНИЯ КАЧЕСТВА ИХ АВТОМАТИЧЕСКОГО ОЗВУЧИВАНИЯ

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

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

Исследованный материал

16305 текстов реальных SMS-сообщений, отправленных в Украине в 2008 году.

Пять категорий SMS-текстов в соответствии с языком, на котором они написаны. К категории «условно-украинский язык» отнесены:

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

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	10825 (100%)	5480 (100%)	

Многоязычность текстов SMS-сообщений

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

SMS-сообщения на украинском и условно-украинском языках

Критерии разграничения SMS-текстов

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

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

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

SMS-сообщения с нестандартной транслитерацией

17% SMS-сообщений написаны с использованием латинского алфавита. В половине случаев игнорируется наличие официальных таблиц транслитерирации.

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.

Транслитерация украинских SMS-сообщений русскими буквами, например, «*А ты де на Днипри?*».

SMS-сообщения с нестандартной орфографией

Нестандартная орфография зафиксирована в 39% украинских SMS (не считая SMS на суржике).

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

SMS-сообщения с нестандартными сокращениями, элементами сленгов и ненормативной лексики

Нестандартные сокращения в SMS-текстах не могут быть расшифрованы.

Молодежный и SMS сленги отличаются динамичностью, необходим мониторинг словарей. Слова из списка ненормативной лексики заменяются сигналом «би-и-п».

Экспрессивный характер SMS-сообщений требует особой выразительной просодики.

SMS-сообщения на суржике

На суржике общаются 15% взрослого населения Украины и 27% студентов. 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.

Выводы

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

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

Для озвучивания SMS, написанных на суржике, необходимо либо расширение речевых баз данных украинского языка, либо создание специальных речевых баз данных суржика.

Образец озвученных SMS на украинском языке

- Чому не пишете, для чого тоді дзвонили? Чого хотіли? Перезвоніть. Яна.
- це хто? напиши смс.
- я тебе дуже люблю!
- Привет. коли ти будеш?
- Оля! Ми вибачаємось. Не їдьте. Вибач.
- будьте щасливі.
- щиро вдячна.
- Котінька, люблю тебе!
- +050 944-80-81.
- +380984929526.

Образец озвученных SMS на украинском языке

- Ви отримали SMS повідомлення.
- Відправник.
- Дата відправки.
- Дурнику, привіт!
- Всім привіт з Полтави.
- До побачення.
- Будь ласка.
- Мамо, передзвони братові.
- Передзвоніть мені.

Образец озвученных SMS на украинском языке

- Будь ласка, зателефонуйте.
- Я вже вдома, подзвони.
- Вибач, що не подзвонив.
- Дзвоніть!
- Зателефоную пізніше.
- Передзвоніть мені, будь ласка.
- Передзвоню пізніше.
- Спасибі, це мій новий номер.
- Терміново зателефонуйте оператору лайф.
- Не доставлено: Невірний номер. Доставлено як голосове повідомлення.

Образец озвученных SMS на украинском языке (общие вопросы)

- а м+ожна Женю?
- Вам Білик давав мій диск з фонограмою?
- Вам Білик дав+ав мій диск з фонограмою?
- Ви ще не там?
- Вийдеш?
- ви тільки на вокзалі? чи вже на місці?
- До тебе можна?
- А можна Інну до телефону?
- А м+ожна Інну до телефону?
- Ви вже вдома?
- Ви сьогодні будете?

Образец озвученных SMS на украинском языке (специальные вопросы)

- Ви хто?
- Ви де?
- Як ви там?
- коли чекати?
- До кого звониш?
- Як настрій?
- Хто це?
- Чого не береш трубку?
- що в тебе нового?
- що робиш? як провела канікули?
- Як у тебе справи? Як життя?

Образец озвученных SMS на украинском языке (интонация завершенности и восклицания)

- Доброго дня. Доброго дня!
- Незабаром літо. Незабаром літо!
- Удачі. Удачі!
- я подзвоню пізніше сама. я подзвоню пізніше сама!
- Вітаю зі святом. Вітаю зі святом!
- приведи друзів. приведи друзів!

Образец озвученных SMS на украинском языке

- вітаю зі святом!
- вітаю з світлим святом Пасхи! Христос воскрес!
- АЛЕ ЦЕ ВЖЕ ЗАНАДТО! ТИ ПЕРЕГИНАЄШ ПАЛКУ!
- ДЯКУЮ! ВИБАЧАЙ!
- Вітаємо з днем народження! Вітаю з днем народження!
- Вітаю з днем банківського працівника!
- Всього самого найкращого!
- Бажаю щастя, здоров'я, сімейного затишку!
- Вітаю з останнім дзвоником!
- Вітаю з першим днем літа!
- Удачі і всього самого тобі найкращого!
- Га, я у Києві!

Образец синтезированной русской речи