УДК 004.912

Ю.В. Крак, А.В. Бармак, Р.А. Багрий, И.О. Стеля

СИСТЕМА ВВОДА ТЕКСТА ДЛЯ АЛЬТЕРНАТИВНОЙ РЕЧЕВОЙ КОММУНИКАЦИИ

Введение

Идея инклюзии — включение людей с ограниченными возможностями жизнедеятельности в общественную жизнь — активно внедряется в современном мире. Более 30 лет как создано направление и самостоятельная область междисциплинарного знания альтернативной и дополнительной коммуникации (Augmentative and Alternative Communication — AAC) [1].

Средства AAC используются, в частности, для предоставления помощи людям, у которых из-за врожденных или приобретенных расстройств отсутствует или существенно ограничена устная речь. При этом дополнительная коммуникация востребована людьми с недостаточно сформированным устным языком и представлена системой специальных методов и средств, призванных, с одной стороны, помочь детям с запаздыванием речевого развития, с другой — облегчить понимание вербальных сообщений лиц с тяжелыми языковыми нарушениями. AAC актуальна в случае отсутствия устного языка и предусматривает овладение абсолютно другой коммуникативной системой, где особое значение приобретают невербальные коммуникативные средства (предметы, фотографии, жесты, символы и т.п.).

AAC использует целый спектр разнообразных способов, которые помогают людям высказывать свои мысли и эффективно общаться. Современные средства AAC представлены тремя основными группами.

1. Простые средства: предметы, жесты, фотографии и изображения, символы. Они помогают выражать потребности в коммуникации, сигнализировать о насущных потребностях жизнедеятельности, получать представление о последовательности событий на протяжении определенного промежутка времени, а также овладевать основными языковыми структурами.

2. Средства коммуникации с использованием простой техники: магнитофоны с проигрыванием одного сообщения, голосовые игрушки, записные книжки и фотоальбомы, коммуникаторы и т.п. С их помощью можно записывать и сохранять в памяти голосовые сообщения и тем самым создавать условия для активного общения людей со значительными ограничениями вербальной коммуникации.

3. Многофункциональные средства коммуникации на основе сложных технических устройств: сенсорные экраны, синтезаторы языка и т.п. Они обеспечивают значительное расширение словаря, позволяют задавать тему беседы и объединять одновременно несколько тем, делают возможным общение на расстоянии, упрощают общение в группе и по телефону.

Современное развитие вычислительной техники и информационных технологий позволяет существенно развить именно третью группу средств AAC. Сущест-

вование в современном обществе групп людей с ограниченными возможностями для общения (с недостатками слуха, с челюстно-лицевыми травмами, после инсультов и т.п.) побуждает к исследованиям альтернативных способов коммуникации.

Авторы предлагают информационную технологию для реализации альтернативных подходов к общению [2]. Основная идея технологии — создание комплексного подхода к реализации альтернативной коммуникации. Аппаратно-программная реализация технологии должна обеспечить коммуникацию максимально возможными способами.

Важный шаг в предложенной информационной технологии — интеллектуализация ввода информации. Цель работы — исследование интеллектуализации ввода информации с помощью системы ускоренного ввода текста в цифровые устройства. Такая система использует меньшее количество команд для ввода букв и прогнозирует варианты слов, базируясь на данных корпуса слов и словосочетаний для общения.

Обзор интеллектуализации и альтернативных способов ввода текста

Обзор способов интеллектуализации ввода текста показал, что указанная проблема возникла и решалась для систем ввода текстовой информации в мобильные устройства с цифровыми кнопками. С помощью девяти цифровых кнопок предлагались различные способы ввода буквенной информации. Имея в виду, что минимизация количества управлений для ввода текстовой информации является важной для альтернативных систем коммуникации, рассмотрим эти способы ввода: Multi-press Input Method, Two-key Input Method и T9 Input Method [3].

Multi-press Input Method — основной метод ввода текста для мобильных телефонов. Он состоит в нажатии каждой кнопки один или несколько раз, для указания входного символа. Multi-press Input Method имеет проблему сегментации, которая возникает при необходимости использовать одну и ту же кнопку, что и для ранее введенного символа. Система должна определить, что нажатие кнопки «относится к подкатегории» предыдущего или нового символа. Это вызывает необходимость создания механизма, который будет определять начало нового символа. Используют два основных пути преодоления этой проблемы: использование тайм-аута (задержки), на протяжении которого нажатие кнопки будет означать тот же символ или наличие соответствующей кнопки, чтобы убрать задержку и ввести сразу следующий символ на той же кнопке. Некоторые модели телефонов используют комбинацию из этих двух решений.

В Two-key Input методе пользователь нажимает на две кнопки последовательно, указывая символ. Первое нажатие выбирает «группу» символов, второе устраняет неоднозначность (для указания положения символа в группе).

Наиболее заслуживающим внимания является T9 Input Method [3], который базируется на запатентованной технологии для мобильных телефонов. Технология разрешает вводить слова в SMS-сообщениях одним нажатием клавиши для каждой буквы (клавиша содержит несколько букв). Таким образом, вместо нажатия кнопки несколько раз (например, чтобы ввести букву «В»), нужно нажать клавишу только один раз. Программное обеспечение попробует предсказать, какую букву нужно выбрать («А, «Б», «В» или «Г»), используя словарь и предлагая наиболее релевантный вариант. Итак, технология разрешает быстро вводить текст на цифровой клавиатуре, а также на небольших сенсорных экранах (например, в смартфонах). Даже для больших сенсорных экранов указанная технология полезна: она позволяет уменьшить размер виртуальной клавиатуры для ввода текста.

T9 Input Method решает проблему неоднозначности предсказания путем использования определенных лингвистических знаний, например словаря как основы для преодоления неоднозначности. Метод базируется на том же расположении

кнопок, что и в Multi-press Input Method, но каждая кнопка нажимается только один раз. T9 находит возможные слова для такой комбинации кнопок по лингвистической базе и «предугадывает» ожидаемое слово. Лингвистическая неоднозначность не является совершенной, так как много простых слов могут иметь одну и ту же последовательность кнопок. В этих случаях T9 предлагает слово, которое чаще используется, как слово по умолчанию. Чтобы выбрать другое слово, нужно нажать соответствующую кнопку.

Приведенный в исследованиях анализ [4] показал, что устранение неоднозначности работает довольно эффективно. Так в выборке из 9025 наиболее употребляемых слов на английском языке (ftp://ftp.itri.bton.ac.uk/), полученных из Британского национального корпуса, неоднозначность присутствует только для 3 % слов.

Альтернативные способы общения, как правило, более медленные, чем разговорные. В системах AAC, которые часто используются (адаптивные клавиатуры, джойстики, сенсорные панели и т.п.) скорость передачи данных часто составляет меньше десяти слов в минуту [5], в сравнении с 150–200 слов в минуту для обычного языка. Если рассматривать более сложные AAC устройства, такие как системы, которые базируются на мозг-компьютерном интерфейсе, то скорость передачи данных снижается меньше, чем до пяти символов в минуту [6]. Для увеличения скорости передачи данных были предложены различные стратегии и подходы, которые заключаются в кодировании слова, фразы или предложения сокращениями, которые автоматически расширяются при вводе текста. Такой подход может быть полезным для очень распространенных слов или фраз, но редко применяется в контекстно свободной коммуникации.

Другие системы используют информацию о частотных характеристиках обычного языка, чтобы менять местами или кодировать символы в удобном виде. Например, AAC устройства, которые используют определенные речевые устойчивые выражения, могут переставлять их таким образом, что чаще используемые становятся первыми, это ускоряет выбор [7].

Более сложные системы AAC используют методы обработки естественного языка (NLP) на базе моделирования и прогнозирование языка в целях повышения скорости ввода [8]. Эти методы либо прогнозируют следующий возможный символ, либо предлагают наиболее возможное слово после предыдущего.

Прогнозы высчитываются совмещением статистических данных, извлеченных из учебных корпусов, с информацией о текущем контексте предложения. Они обычно базируются на частотных таблицах, синтаксической структуре, семантической информации или их комбинации.

Методы, которые базируются на прогнозе слов, наиболее часто используются в системах AAC, в них список возможных слов обычно предлагается для выбора. Список прогнозирования может предлагаться отдельно или появляться в пределах текстового поля для ввода, или на экране пользовательского интерфейса.

Очевидно, что расширение системы AAC функцией прогнозирования может как увеличить скорость передачи, так и вызвать проблемы в интерфейсе пользователя и во взаимодействии с ним. Для оценивания реальной пользы предсказаний должны быть приняты во внимание многие факторы: фактическая размерность списков возможных слов, соображение о задержках и когнитивных усилиях и т.п. [9, 10].

В данной работе предлагается использовать методы прогнозирования, которые базируются на моделях, построенных на базе корпусов фраз украинского языка, свойственных конкретным жизненным ситуациям.

Исследование системы ввода текста с прогнозированием

Как уже отмечалось, основное назначение технологий, основанных на алгоритме T9 Input Method, состоит в возможности прогнозирования следующих слов и фраз при коммуникации. Алгоритм прогнозирования способен автоматически завершать ввод текста, что позволяет оптимизировать время ввода. Для прогнозирования нужно найти баланс между скоростью ввода и функциональностью. Чем больший словарь и чем больше технологий ввода используется (N-grams, нечеткий поиск, фонетический поиск), тем медленнее становится время отклика, но повышается качество результатов.

Для исследования и проектирования системы ввода текста с прогнозированием предлагается следующая модель, представленная соответствующими параметрами:

- множество (корпус) слов украинского языка (ограниченный словами для повседневного общения);
- множество букв украинского алфавита в определенном порядке следования: алфавитный, клавиатурный (QWERTY), по частоте использования и т.п.;
- группирование букв (количество групп и количество букв в группе).

Рассмотрим модель, представленную алгоритамами:

- формирование корпуса слов и словосочетаний;
- алгоритм кодирования слов;
- алгоритм прогнозирования.

Для формирования корпуса слов предлагается использовать экспертный подход. Это обусловлено тем, что нужно подобрать слова и словосочетания из повседневного общения. Для этого используются источники из контента украиноязычных сайтов, периодической печати, словарей-разговорников и т.п. Полученные таким образом предложения предлагается разбить на *N*-граммы: юниграммы, биграммы и триграммы [11] (рис. 1).

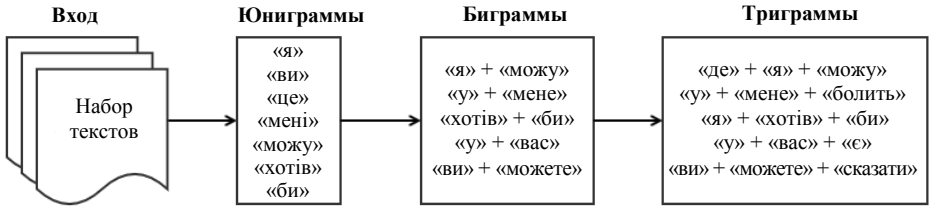

Рис. 1

При обработке естественного языка *N*-граммы используются в основном для предсказания на базе вероятностных моделей. В *N*-граммной модели рассчитывается вероятность последнего слова *N*-граммы, если известны все предыдущие. При этом подходе для моделирования языка предполагается, что появление каждого слова зависит только от предыдущих слов.

Цель построения *N*-граммных моделей — определение вероятности использования заданной фразы (словосочетания). Эту вероятность можно задать формально как вероятность возникновения последовательности слов в некотором корпусе (наборе текстов). Для оценки этих вероятностей нужен соответствующий метод. Самый простой и наиболее интуитивный способ оценки достоверности — метод максимального правдоподобия (Maximum Likelihood Estimation (MLE)) [11].

Для оценки параметра максимального подобия слов нужно определить параметры, которые максимизируют достоверность этого сходства для заданных слов. MLE-оценка параметров модели *N*-граммы может быть получена как нормализованное количество из корпуса. Корпусом является набор текстов, который стати-

стически репрезентативен для моделирования языка. Например, можно оценить биграмм-вероятность слова w_n, учитывая предыдущее слово w_{n-1}, подсчитывая вхождение биграмм $C(w_n, w_{n-1})$ и нормируя суммой всех биграмм, содержащих первое слово w_n :

$$P(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n)}{\sum C(w_{n-1}, w)}. \quad (1)$$

Поскольку количество всех биграмм, начинающихся со слова w_{n-1}, равно количеству юниграмм для этого слова w_{n-1}, выражение (1) упрощается:

$$P(w_n \mid w_{n-1}) = \frac{C(w_{n-1}, w_n)}{C(w_{n-1})}. \quad (2)$$

В общем случае *N*-грамм модели формула для оценки параметров MLE будет

$$P(w_n \mid w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}, w_n)}{C(w_{n-N+1}^{n-1})}. \quad (3)$$

Модели *N*-грамм в основном используются для предсказания слова, так как эффективны и просты в использовании. К тому же статистика частот, которая используется для расчета MLE оценки, может быть получена непосредственно из текстов без особых усилий.

Для реализации T9 Input Method нужно предложить распределение букв по группам и указать порядок следования их в этих группах. Распределение букв по группам является задачей оптимизации, так как от него зависит количество слов с одинаковым кодом и, как следствие, скорость работы пользователя с программой. Для определения оптимального распределения нужно проанализировать предложенные порядки следования букв и определить критерии, которые влияют на конечный результат.

Алфавитный и клавиатурный (QWERTY) порядок следования букв общеизвестен, и поэтому их следует рассмотреть в задаче распределения на группы. На рис. 2 изображен пример распределения на шесть групп для обоих случаев.

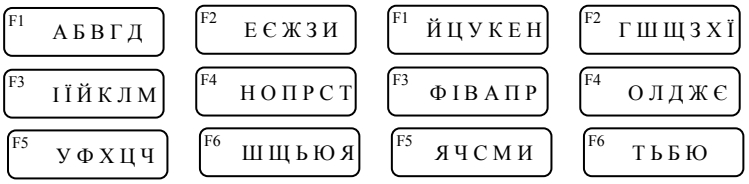

Рис. 2

Для обоих этих случаев нет возможности изменять порядок их следования, поэтому можно провести только исследование с неравномерным распределением букв по группам и определить наилучшие варианты.

Другой подход — использование частотного порядка следования букв. Для украинского языка есть много литературных источников, где приведены среднестатистические частоты повторяемости букв и их биграммы (двухбуквенные последовательности). В работе [12] приведены среднестатистические частоты букв без учета пропуска между словами на основе анализа страниц украиноязычных сайтов и разных стилей современного украинского языка, в том числе разговорно-бытового. На рис. 3 приведен порядок следования букв согласно частоте использования.

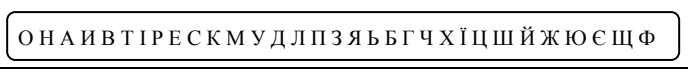

Рис. 3

Для поиска оптимальных распределений букв условно разделим множество букв на классы по частоте их использования. В пределах этих классов будем изменять порядок следования букв, придерживаясь распределения букв с близкими частотами по разным группам. Исходя из этих требований, предложен алгоритм такого распределения. Пример работы алгоритма для случайного распределения частотного порядка следования букв на четыре группы приведен на рис. 4.

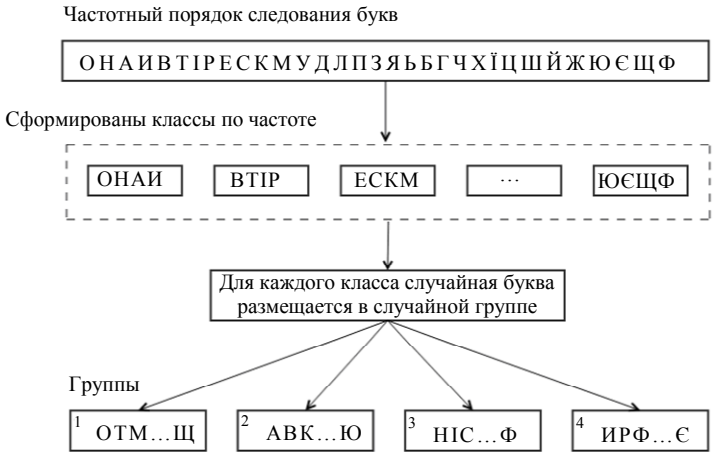

Рис. 4

Алгоритм работает следующим образом:

1) буквы в частотном порядке следования разбиваются на восемь классов таким образом, чтобы в одном классе были буквы с близкими значениями частот;

2) для каждого класса выбирается случайная буква и размещается в группе, где еще нет буквы из этого класса.

Полученные таким образом группы сохраняют заданное требование, т.е. не допускают нахождения в одной группе букв с близкими частотами.

Следующий анализ среднестатистических частот повторяемости биграмм в украинском языке [12] показал, что для него характерно чередование гласных и согласных. Поэтому для нахождения лучшего распределения следует отделить гласные и согласные и сформировать две частотные последовательности следования букв — для 12 гласных и 20 согласных (рис. 5).

Рис. 5

Алгоритм формирования букв аналогичный предыдущему, кроме того, что классы по частоте букв формируются отдельно для гласных и согласных. Это позволяет сформировать такие группы, где кроме учета частоты буквы, еще обеспечивается уникальность кода для наиболее частотных биграмм. Так, для четырех групп 16 наиболее частотных биграмм будут иметь уникальные коды, что позволяет значительно уменьшить количество слов с одинаковым кодом.

Все слова корпуса кодируются для заданного расположения символов по группам. Такое кодирование необратимо. Код является набором цифр, которые соответствуют группам, в которых находится каждая буква слова. Например, для слова «хочу» и стандартного расположения букв на цифровой клавиатуре телефона код будет 7576. Декодирование сводится к поиску в словаре кода (полученного на входе) и возвращению соответствующего этому коду слова. Алгоритм кодирования приведен на рис. 6.

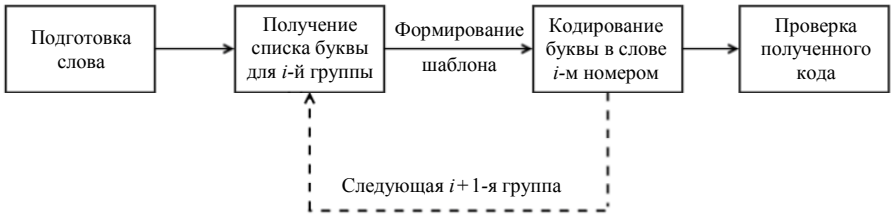

Рис. 6

Пример кодирования слова «болить» для клавиатурного распределения букв на шесть групп приведен на рис. 7.

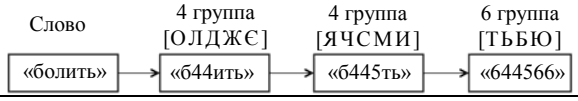

Рис. 7

Для исследования лучшего варианта распределения на группы предложен алгоритм определения количества слов с одинаковыми кодами.

Для каждого слова из таблицы юниграмм (корпуса слов) выполняем поиск слов, которые имеют действительный код, группируем список слов по коду, отбираем все коды, которые имеют количество повторов больше единицы, суммируем общее количество одинаковых кодов.

Основные экспериментальные результаты

Проведена серия тестов для разных порядков следования букв и со случайным распределением по группам.

Количество одинаковых кодов для разного количества групп для алфавитного и клавиатурного порядка следования буквы приведены в табл. 1.

Таблица 1

Последовательность букв	Количество слов с одинаковым кодом				
	4	5	6	7	8
алфавитная	807	518	365	273	227
клавиатурная	714	487	307	220	96

Для всех групп распределение клавиатурного порядка следования букв показало лучшие результаты, чем алфавитного. Это связано с тем, что расположение букв на клавиатуре соответствует частоте использования букв в словах — более популярные в центре, менее популярные по краям. Также с увеличением количества групп количество слов с одинаковым кодом линейно уменьшается.

Следующий тест предполагал случайное формирование групп, но с учетом частоты букв (частотная) и частоты биграмм (частотная (2)). В табл. 2 для сравнения дополнительно приведены результаты для алфавитной и клавиатурной последовательности букв, хотя случайное формирование групп для этих последовательностей не имеет смысла с точки зрения их природы.

Таблица 2

Последовательность букв	Количество слов с одинаковым кодом (лучшая итерация / в среднем)				
	4	5	6	7	8
алфавитная	538 / 751	301 / 508	168 / 355	112 / 274	73 / 216
клавиатурная	545 / 756	325 / 499	175 / 364	123 / 280	87 / 222
частотная	532 / 701	306 / 448	176 / 324	94 / 219	65 / 176
частотная (2)	515 / 626	260 / 373	137 / 199	88 / 145	56 / 103

Из проведенных исследований получено, что частотная последовательность следования букв с учетом гласных/согласных дает лучшие результаты распределения в среднем и быстрее находит минимальные варианты. Для этой последовательности отличие от других становится наиболее заметным с увеличением количества групп. Это связано с тем, что количество возможных уникальных кодов для биграмм увеличивается вплоть до 64 (для восьми групп), а значит, большинство слов получат уникальные коды. На рис. 8 приведены наиболее удачные распределения для частотной (2) последовательности букв для разного количества групп.

Рис. 8

На рис. 9 изображен график, который показывает отношение количества слов с одинаковым кодом к общему количеству слов в словаре для распределения букв на группы разных последовательностей следования букв.

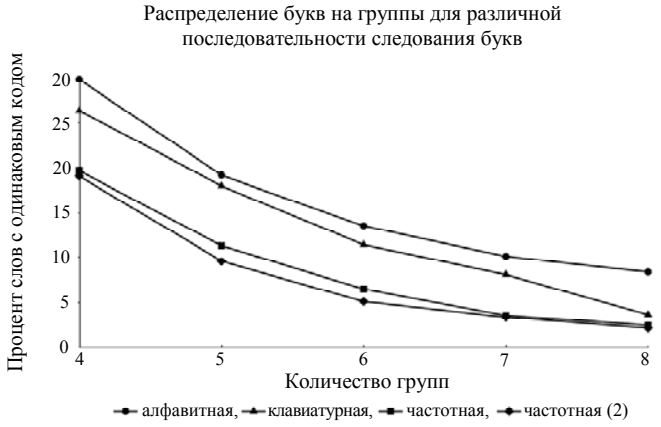

Рис. 9

Проанализировав данные, приведенные на рис. 9, можно сделать следующие выводы.

1. Для восьми групп (подобное распределение используется в мобильных устройствах) количество слов с одинаковым кодом для всех последовательностей находится в пределах до 10 % от общего количества слов в словаре. Следовательно, целесообразно использовать наиболее удобную для пользователя последовательность следования букв — алфавитную или подобную клавиатуре (рис. 10, 11). Такие порядки следования букв общеизвестны и не требуют много времени на адаптацию.

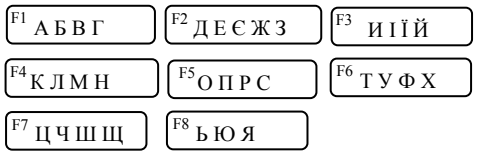

Рис. 10

Рис. 11

2. Для шести групп в пределах 10 % остаются только частотные последовательности следования букв. Для других последовательностей количество слов с одинаковым кодом значительно увеличивается, и их использование становится непрактичным. На рис. 12 показано наилучшее распределение для частотной последовательности следования гласных и согласных (частотная (2)). Количество слов с одинаковым кодом для данного распределения составляет 5,1 % от общего количества.

Рис. 12

3. Для четырех групп количество одинаковых кодов для всех последовательностей следования букв становится больше 20 %, и для комфортной работы необходимо использование дополнительных алгоритмов прогнозирования для определения наиболее ожидаемого слова.

Заключение

В настоящей работе для имплементации информационной технологии, которая реализует альтернативные подходы к общению [2], предложен подход к интеллектуализации ввода текстовой информации на основе T9 Input Method [3]. Для украинского языка для реализации предложенного метода ввода текстовой информации исследованы варианты разбивки буквенного ряда на группы. Предложены распределения как для клавиатурного порядка следования букв, так и для порядка, который базируется на частоте использования букв для слов корпуса языка.

Дальнейшие исследования направлены на анализ систем прогнозирования окончаний слов и словосочетаний при использовании T9 Input Method. Полученные результаты войдут в аппаратно-программное обеспечение системы альтернативной коммуникации для людей с отсутствующим каналом основной вербальной коммуникации.

Ю.В. Крак, О.В. Бармак, Р.О. Багрій, І.О. Стеля

СИСТЕМА ВВЕДЕННЯ ТЕКСТУ ДЛЯ АЛЬТЕРНАТИВНОЇ МОВНОЇ КОМУНІКАЦІЇ

Запропоновано підхід до інтелектуалізації введення текстової інформації на основі T9 Input Method. Для реалізації запропонованого методу введення текстової інформації для української мови досліджено варіанти розподілення буквеного ряду на групи. Запропоновано розподілення як для клавіатурного порядку слідування букв, так і для порядку, який базується на частоті використання букв для слів корпусу мови.

Yu.V. Krak, A.B. Barmak, R.A. Bagriy, I.O. Stelya

TEXT ENTRY SYSTEM FOR ALTERNATIVE SPEECH COMMUNICATIONS

Intellectualization of text entry on the basis of T9 Input Method is proposed. For the implementation of the proposed method of text entry for the Ukrainian language, a number of options for the distribution of letters into groups are studied. The distribution for the keyboard-order of letters, and for the order, which is based on the frequency of use of letters for words corps language are proposed.

1. *Augmentative* and Alternative Communication (AAC). — http://www.asha.org/public/speech/disorders/AAC/.
2. *Кривонос Ю.Г., Крак Ю.В., Бармак А.В., Багрий Р.А.* Новые средства альтернативной коммуникации для людей с ограниченными возможностями // Кибернетика и системный анализ. — 2016. — **52**, № 5. — P. 3–13.
3. *Grover D.L., King M.T., Kuschler C.A.* Patent No. US5818437, Reduced keyboard disambiguating computer. Tegic Communications, Inc., Seattle, WA (1998).
4. *Silfverberg M., MacKenzie I.S., Korhonen P.* Predicting text entry speed on mobile phones // Proceedings of the ACM Conference on Human Factors in Computing Systems — CHI 2000. — 2000. — New York : ACM. — P. 9–16.
5. *Newell A., Langer S., Hickey M.* The role of natural language processing in alternative and augmentative communication // Natural Language Engineering. — 1998. — **4**(01). — P. 1–16.
6. *The* brain-computer interface presents the novel mental typewriter hex-o-spell // Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course / B. Blankertz, G. Dornhege, M. Krauledat, M. Schroder, J. Williamson, R. Murray-Smith, K.R. Muller. — 2006. — P. 108–109,
7. *Lesher G., Moulton B., Higginbotham D.J.* Techniques for augmenting scanning communication // Augmentative and Alternative Communication. — 1998. — **14**(2). — P. 81–101.
8. *Copestake A.* Augmented and alternative NLP techniques for augmentative and alternative communication Proceedings of the ACL workshop on Natural Language Processing for Communication Aids. — 1997. — P. 37–42.
9. *The Effects* of word prediction on communication rate for AAC. K. Trnka, D. Yarrington, J. McCaw, K.F. McCoy, C. Pennington, I. AgoraNet / NAACL-HLT; Companion Volume: Short Papers, 2007. — P. 173–176.
10. *Garay-Vitoria N., Abascal J.* Text prediction systems: a survey // Universal Access in the Information Society. — 2006. — **4**(3). — P. 188–203.
11. *Jurafsky D., James H. Martin* Speech and language processing. Copyright c 2014. All rights reserved. Draft of January 9, 2015.
12. *Сушка С.О.* Частоты повторяемости букв и биграмм в открытых текстах на украинском языке. — http://jrnl.nau.edu.ua/index.php/ZI/article/view/1968

Получено 15.09.2016

Статья представлена к публикации членом редколлегии академиком НАН Украины Ю.Г. Кривоносом.