Луканин А.В. Искусственная нейронная сеть для генерации форм прошедшего времени в современном английском языке // Научная сессия НИЯУ МИФИ - 2010. XII Всероссийская научно-техническая конференция "Нейроинформатика-2010": Сборник научных трудов. В 2-х частях. Ч.2. М.: НИЯУ МИФИ, 2010. С. 287-297.

А. В. ЛУКАНИН

ЮУрГУ, Челябинск
ice_lc@mail.ru

ИСКУССТВЕННАЯ НЕЙРОННАЯ СЕТЬ ДЛЯ ГЕНЕРАЦИИ ФОРМ ПРОШЕДШЕГО ВРЕМЕНИ В СОВРЕМЕННОМ АНГЛИЙСКОМ ЯЗЫКЕ

В данной статье предлагается использовать многослойный персептрон для ассоциации орфографической записи основ английских глаголов с аддитивными правилами, преобразующими их в формы прошедшего времени, с целью непосредственного её использования в задачах автоматической обработки текстов. Сеть обучается большому лексикону правильных и неправильных глаголов, после чего тестируется её способность к обобщению на других глаголах. В заключение приводятся рекомендации по улучшению процесса обучения.

Введение

Один из взглядов на язык, восходящий к работам Н. Хомского [1-2], поддерживаемый Дж. Фодором [3] и в данный момент развиваемый С. Пинкером [4-5], предполагает, что абстрактные правила играют ключевую роль в обработке естественного языка. Альтернативная позиция принадлежит так называемому коннекционистскому направлению [6], которое рассматривает все когнитивные процессы, включая обучение языку, как постепенную подстройку весовых коэффициентов связей простых вычислительных устройств — нейронов, совокупность которых составляют искусственные нейронные сети (ИНС) — модели нервной системы человека. Согласно этому взгляду в языке отсутствуют явные правила, но постепенное обобщение фактов использования языка в различных ситуациях, в конечном счёте, настраивает нейронные связи таким образом, что языковая способность человека напоминает систему, основанную на правилах.

Коннекционизм рассматривает различные аспекты языка, но одним из первых вопросов, поднятых в работе [7] и до сих пор не до конца разрешённым, является моделирование прошедшего времени в английском языке. Большинство глаголов в английском языке являются правильными (ПГ): для образования формы прошедшего времени к основе глагола добавляется морфема /d/, которая в зависимости от последней фонемы осно-

вы имеет аллофоны [t], [d] или [ɪd], например, в формах *cooked*, *boiled* и *generated* соответственно. Кроме того, в современном английском языке есть небольшая группа неправильных глаголов (НПГ, по разным оценкам от 120 до 200, устаревшие глаголы во внимание не берутся), для которых форма прошедшего времени образуется по-другому.

С. Пинкер [5] полагает, что английский язык обладает двумя механизмами образования форм прошедшего времени, и в мозге человека есть 2 различных отдела, осуществляющие их: глагол первоначально ищется в лексиконе (в котором содержатся неправильные глаголы) и если слово не найдено в нём, к глаголу применяются правила словоизменения.

Однако формы прошедшего времени неправильных глаголов не являются случайными образованиями: начала обеих форм обычно совпадают (есть только 2 супплетивные формы: *went* для *go* и *was* для *be*). Большинство форм неправильных глаголов заканчиваются, как и формы правильных глаголов на [t] или [d] (59% из 181 глаголов, приведённых в [8]). Основное же отличие заключается в том, что у неправильных глаголов происходит изменение гласной в корне. При этом неправильные глаголы можно подразделить на несколько групп в зависимости от характера изменений. К примеру, в одну группу входят глаголы *keep*, *sleep*, *creep*, *sweep* и *weep*, к формам прошедшего времени которых добавляется аллофон [t], а корневая гласная [i:] сокращается до [e]. Таким образом, неправильные глаголы являются на самом деле псевдо-правильными, так как повторяют большинство фонетических фактов образования форм правильных глаголов.

Дж. МакКлелланд [9] показывает, что искусственные нейронные сети адекватно моделируют основные характеристики познавательной деятельности человека, в то время как модели, основанные на правилах и исключениях, не могут этого продемонстрировать.

После продолжительной критики коннекционистского подхода предлагались различные модификации модели, описанной в [7]. Так, одна из последних успешных моделей описана в [10]. К. Планкетт и П. Хуола используют модель многослойного персептрона для моделирования словоизменения прошедшего времени глаголов и множественного числа существительных. ИНС обучается словоизменению 946 односложных глаголов, 122 из которых являются неправильными. Исследователи сравнивают 2 метода обучения: в первом эксперименте сети показываются все слова, во втором – слова вводятся постепенно, увеличивая размер обучающей выборки на 5%. При обучении используется частота словоформ, взятая из корпуса Брауна и выровненная в диапазоне от 1 до 20. Эффективность

обучения для всех глаголов составляет 99.2% и 80% для неизвестных для ИНС слов (из 1541 односложных существительных и глаголов, количество глаголов не приводится).

Как и предыдущие модели, рассмотренная ИНС моделировала ассоциацию фонетической формы основы с фонетической формой прошедшего времени глагола, исходя из предположения, что именно фонологические факты влияют на словоизменение при овладении языком ребёнком. Однако для того, чтобы использовать такую модель при автоматической обработке текста, необходимо привести орфографическую запись к фонетической, а после получения фонетической формы на выходе ИНС преобразовать её в соответствующую орфографическую запись.

Постановка задачи и описание алгоритма

Как известно, орфографическая система английского языка является довольно сложной, в связи с тем, что множество заимствованных слов сохраняют не только заимствованное произношение, но и орфографию: некоторые буквы не читаются, некоторые читаются по-разному в зависимости от контекста.

В отличие от предыдущих авторов, мы моделируем ассоциацию орфографической записи основы глагола с правилами преобразования её в форму прошедшего времени с целью непосредственного использования данной модели для генерации орфографической записи формы прошедшего времени из соответствующей основы. При этом сами правила в ИНС не прописываются, фактически ИНС обучается относить то или иное слово к определённому классу. Сами же правила могут быть аналогичным образом представлены в виде ИНС, которая будет ассоциировать одну последовательность букв с другой, но эта задача в данном исследовании не стояла.

Мы использовали модель многослойного персептрона с алгоритмом обратного распространения ошибки. Первоначально на вход подавался вектор из 162 бинарных чисел, кодирующих 6 последних букв основы глагола (если длина глагола была меньше 6 букв, в начало добавлялись пробелы). Каждая буква представлялась подвектором из 27 чисел соответствующих 26 буквам английского алфавита и пробелу (число, соответствующее номеру буквы в алфавите, кодировалось единицей, все остальные числа подвектора равнялись нулю). Но такая система кодирования оказалась эффективной только для обучения малому количеству глаголов. Поэтому в последующих экспериментах было решено разделить буквы на гласные и согласные, кодируя их методом «двуполярного термометра».

Окно было расширено до 8 последних букв глагола (в обучающей выборке максимальная длина глагола равнялась 12 буквам), а размер подвектора сокращён до количества согласных — 21. Гласные кодировались «левосторонним термометром» («a» как «10000…», «e» как «11000…», «i» как «11100…» и т.д. к концу алфавита), согласные — «правосторонним термометром» («z» как «…00001», «y» как «…00011», «x» как «…00111» и т.д. к началу алфавита, т.е. подвектор для буквы «b» состоял только из единиц). Конфигурация ИНС: 168x20x66, темп обучения: 0.1.

Для экспериментов мы взяли глаголы из лексикона английского языка Englex морфологического анализатора PC-Kimmo [11] за исключением устаревших слов (всего 3591 глагол). Нами рассматривался американский вариант английского языка, поэтому, сгенерировав в PC-Kimmo различные формы глаголов для британского и американского вариантов, в MS Word 2003 с выбранным модулем проверки орфографии для американского варианта английского языка мы удалили все британские варианты написания форм прошедшего времени (например, «disembowelled» для американского «disemboweled»).

Проанализировав различия между основой и формой прошедшего времени, нами было выделено 85 классов словоизменения, которые были сведены к 66 аддитивным правилам: 4 нейрона в выходном слое отвечают за удаление от 1 до 4 букв с конца, 62 других нейрона кодировали добавление к концу слова 62 разных комбинации букв (16 из них используются для кодирования словоизменения правильных глаголов: 13 разных вариантов удвоения последней согласной, добавление *ied*, *d* и *ed*).

Каждой форме прошедшего времени была поставлена в соответствие частота из корпуса Брауна [12], при этом несколько устаревших неправильных глаголов были заменены современными правильными глаголами (к примеру, в корпусе найдена форма «learned», но отсутствует форма «learnt»). Все найденные в корпусе формы были включены в обучающую выборку (2323 глагола), кроме того, в обучающую выборку были добавлены неправильные глаголы *beseech, slink, smite, tread* и *wring* с частотой 1. Остальные 1263 глагола (79 из которых – неправильные) использовались для тестирования ИНС. Первоначально частота глаголов была выровнена по формуле $\log_3(F_{гл}+1)$ так, чтобы самый частотный глагол был в 9 раз чаще наименее частотного, но наименее частотные неправильные глаголы не заучивались сетью, поэтому для неправильных глаголов была внесена поправка и их частота рассчитывалась по формуле:

$$F_{непр.гл.} = \log_3\left(F_{непр.гл.}+1\right) + \frac{\log_3\left(F_{\max}+1\right) - \log_3\left(F_{непр.гл.}+1\right)}{2} \qquad (1)$$

т.е., если частота правильных глаголов лежит в диапазоне [1,9], то частота неправильных глаголов — в диапазоне [5,9], при этом все 170 неправильных глаголов находились в начале списка среди первых 272 глаголов. Первоначальная выборка содержала первые 20 глаголов и через каждые 5 эпох увеличивалась на 5%, т.о. вся обучающая выборка была показана через 109 приращений на 546 эпоху (3-я вертикальная разделительная черта на рис. 1-2), последующие приращения прекращались. Оценка предсказания относительно всей обучающей выборки производилась в период приращений каждую 5 эпоху и, начиная с 546 эпохи, каждую эпоху. Первый правильный глагол при этом был показан на 176 эпоху (1-я вертикальная разделительная черта на рис. 1-2).

ИНС обучалась в течение 1100 эпох, словоизменение неправильных глаголов было выучено на 96.5% (164), правильных – на 98.7% (2129). При тестировании правильно было сгенерировано 85.9% (1017) правильных глаголов и 6% (5) неправильных глаголов (всего 80.9%), что говорит о том, что ИНС успешно обобщила опыт словоизменения правильных глаголов, а неправильные глаголы в основном запомнила как исключения.

Анализ этапа обучения

При анализе процесса обучения необходимо разграничивать динамику обучения (рис. 1) и динамику предсказания глаголов, которые попадут в обучающую выборку (рис. 2).

Как видно на рис. 1 ИНС сначала обучалась довольно эффективно, достигнув максимума на 109 эпоху и выучив 40 из 42 показанных неправильных глаголов, затем количество выученных глаголов в отношении показанных начало снижаться. Это объясняется тем, что до 106 эпохи сети предъявлялось только по 1 глаголу каждые 5 эпох. Начиная со 106 эпохи количество увеличивалось уже на 2 новых глагола, поэтому ИНС стала «не успевать» перестраивать весовые коэффициенты. Со 176-й по 236-ю эпоху падение становится ещё круче, что связано с одной стороны с увеличением количества предъявляемых неправильных глаголов (инкремент достигает уже 5 глаголов), а с другой стороны, сети начинают предъявляться правильные глаголы. Второй фактор ещё больше затрагивает весовые коэффициенты, т.к. входные данные требуется приписывать 2 новым классам – активировать нейрон добавления буквы *d* или нейрон добавления сочетания *ed*.

Набрав критическую массу, на 240-ю эпоху ИНС начинает распознавать правильные глаголы. Однако вплоть до 301 эпохи график количества выученных правильных глаголов очень изрезан, что связано с вводом но-

вых аддитивных правил (*y>ied* и удвоение последней согласной) и показом последних неправильных глаголов из обучающей выборки. Среди последних есть, к примеру, глагол *tread*, требующий отдельного аддитивного правила «добавить *od*» (также активируются нейроны уже выученных правил удаления 3-х букв с конца). С каждым новым правилом темп обучения снижается. К примеру, заметно, что темп обучения правильным глаголам увеличился с 301 по 320 эпоху почти на 30%, т.к. в этот период новых правил не появлялось, и снова замедлился на 321 эпоху в связи с появлением глагола *occur*, который необходимо ассоциировать с нейроном, активирующим новое правило «добавить *red*».

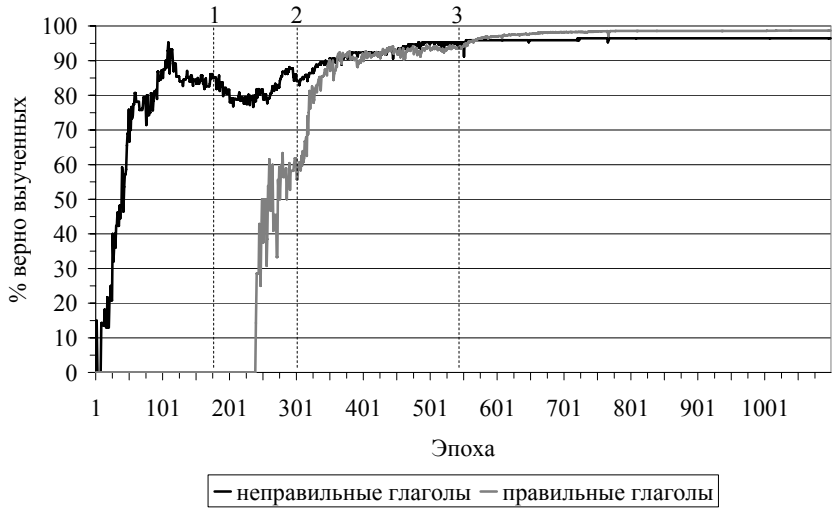

Рис. 1. Процесс обучения ИНС относительно количества показанных глаголов обучающей выборки. 1-я вертикальная граница показывает начало показа первого правильного глагола, 2-я — конец показа всех неправильных глаголов, 3-я — конец показа всех правильных глаголов.

После показа всех неправильных глаголов (2-я вертикальная граница на рис. 1) происходит донастройка весовых коэффициентов, чему также способствуют новые правильные глаголы. Как не удивительно, процесс обучения неправильным глаголам практически останавливается после показа всех правильных глаголов. Видимо, значения весовых коэффициентов находятся в точках локального минимума и могут выйти из них

только благодаря новым входным сигналам, «встряхивающим» нейронную сеть. На рис. 2 также видно, что обучение правильным глаголам (ПГ) происходит эффективнее, чем неправильным (НПГ), однако после показа всех глаголов крутой подъём практически вырождается в горизонтальную линию.

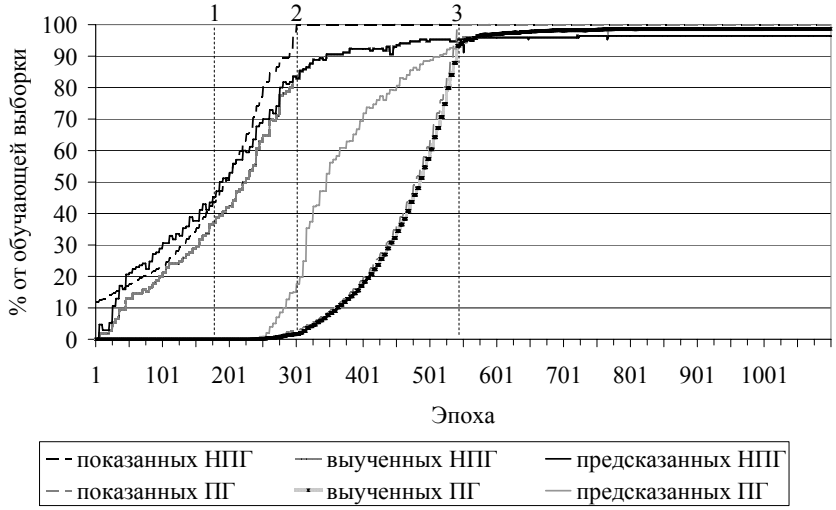

Рис. 2. Процесс обучения ИНС относительно всех глаголов обучающей выборки. 1-я вертикальная граница показывает начало показа первого правильного глагола, 2-я — конец показа всех неправильных глаголов, 3-я — конец показа всех правильных глаголов.

Во время обучения 6 из 170 (3.5%) неправильных глаголов не были выучены полностью, при предъявлении этих глаголов сети активируются только нейроны, соответствующие правилам удаления: *build>*buil, fight>*f, fly>*fl, wake>*w, awake>*aw, dwell>*dwel*[1]. Соответственно, для этих глаголов нужно сделать дополнительную поправку, разместив их выше по списку. Интересно заметить, что первые 5 глаголов очень часто встречаются в разговорах с маленькими детьми («давай построим домик»,

[1] Несуществующие в языке формы принято обозначать в лингвистике символом *.

«не дерись», «птичка улетела», «просыпайся», «он не спит» и т.д.), однако в письменной речи (судя по корпусу Брауна) они представлены реже.

Из 2158 правильных глаголов не было выучено только 29 глаголов (1.4%), причём 23-м из них не было приписано ни одного аддитивного правила (**occur, *grin, *blur, *ban, *comb, *quarrel, *cram, *audit, *panic, *row, *whiz, *abhor, *befog, *brim, *char, *moo, *picnic, *reap, *rev, *stave, *traffic, *trek, *woo*), т.е. они заучились как неправильные глаголы. У одного глагола не была удалена последняя буква перед добавлением *ied* (**occupyied*). Глаголов с правилом *y>ied* среди правильных в корпусе Брауна меньше всего (110) и только в этом правиле необходимо удалить последнюю букву, формы прошедшего времени для остальных правильных глаголов формируются только путём добавления букв. Следовательно, можно заключить, что с точки зрения орфографии такие глаголы являются псевдо-правильными и для них также необходимо делать поправку, размещая выше по списку обучающей выборки.

Глаголу *transfer* кроме правильного добавления *red* было приписано и добавление *ed*, такие ошибки могут делаться детьми и взрослыми в спонтанной речи. Форма **zerod* вместо *zeroed* появилась как следствие сделанного нейронной сетью обобщения, что после гласных обычно идёт окончание *d* вместо *ed*. В обучающей выборке был только один похожий глагол (*echo>echoed*), который стоял в середине списка, что помогло нейронной сети запомнить это исключение.

Кроме того, 3-м глаголам было приписано правило добавления *ed* вместо *ted* (**transmited, *acquited, *manumited*). Это объясняется тем, что наша модель не располагает информацией о месте ударения. При стандартном прочтении английских глаголов ударение ставится на первый слог (например, *limited, visited*) или второй, если присутствует приставка (например, *inherit*), поэтому удвоение последней согласной не требуется. В односложных словах удвоение также не требуется, если гласный звук является сложным (*waited, suited*), запрещая произнесение *i* в данной позиции как [aɪ] (как это было бы для *like>liked*). Произношение этих глаголов требуется запоминать, что может быть также решено введением соответствующей поправки.

Предложенный порядок обучения ИНС английским глаголам подтверждает факты запоминания глаголов детьми и может быть также использован при разработке учебников и методических пособий по обучению английскому языку школьников и студентов.

Анализ этапа тестирования

Нейронная сеть обучается предсказывать новые неправильных глаголы, однако количество НПГ в обучающей выборке не столь велико, чтобы вывести общие правила. На 170 глаголов приходится 46 аддитивных правил (в сочетании с 4 правилами удаления букв, которые нейронная сеть разделяет с правильными глаголами), плюс 1 правило «не менять форму», выражающееся в нулевых значениях всех выходных нейронов. Видимо, поэтому ИНС плохо обобщает полученные данные по неправильным глаголам, заучивая их.

При тестировании ИНС было показано 1263 глагола (79 неправильных и 1184 правильных), из которых правильно было сгенерировано 6% (5) неправильных глаголов и 85.9% (1017) правильных глаголов (всего 80.9%).

Ошибки в генерации форм прошедшего времени НПГ можно разделить на 5 классов:

- 54 глагола образованы по правилам образования форм ПГ (**backbited, *housekeeped, *merrymaked, *slayed, *wrongdod* и др.);
- 7 глаголов не изменили форму (**babysit, *betake, *bookkeep, *bricklay, *firefight, *frostbite, *shipbuild*);
- у 1 глагола были только удалены буквы (*gird>*g*);
- для 4 глаголов были выбраны неверные правила (**begetreded, *bookbindted, booksell>*booksede, uprise>*uprid*);
- 8 глаголам были сопоставлено более 1 правила (**cabinetmakeded, *foretellolded, *gunsliunged, *homebreakded, *housebreakded, *inbreedded, *lawgaved, *strikebreakded*).

Все предъявляемые неправильные глаголы кроме глагола *gird* являются сложными глаголами, образованными из простых глаголов добавлением приставки или корня другого слова. Таким образом, мы полагали, что ИНС научится определять, что если простой глагол совпадает с концом сложного глагола, то для последнего активировать нужно те же выходные нейроны, что и для простого глагола. Однако чтобы сеть сделала такое обобщение, необходимо было предъявлять ей несколько примеров для каждого набора активирующихся аддитивных правил. Из 170 неправильных глаголов обучающей выборки сложных было 23. Из 6 правильно сгенерированных форм прошедшего времени НПГ 2 не требовали изменений (*browbeat, telecast*), что является очень продуктивным правилом (30 из 170 показанных глаголов), но 4 глагола, требующие это правило, были сгенерированы неверно: **leapreaded, *proofreaded, *simulcasted, *typeseted*.

Введённая нами кодировка слов чётко разграничивает гласные, согласные и пробел. В виду малого количества НПГ кодировка слов и размер окна оказывают большое влияние на процесс обобщения. Так, форма *telecast* была сгенерирована верно, т.к. ИНС сделала обобщение по структуре известных ей глаголов *broadcast* и *forecast*: CVVCcast и CVCVcast (где C – любая согласная, V – любая гласная). Это подтверждают результаты тестирования ИНС по дополнительному лексикону искусственно созданных глаголов: глаголы *voicecast, corecast, carecast, datacast, coatcast, baitcast* не меняют свою форму, глаголы *precast, recast, lampcast, lightcast* активируют нейрон добавления *ed*.

Таким образом, мы полагаем, что для более эффективного обучения ИНС в обучающую выборку необходимо также включать искусственно созданные сложные неправильные глаголы, где для каждого простого НПГ в его составе будет присутствовать слова с разным составом согласных, гласных и разным количеством пробелов в начале слова в рамках окна входных данных. Отграничив таким образом модели словоизменения НПГ друг от друга и от моделей ПГ, ИНС, построенная по предлагаемой схеме, должна эффективнее обобщать данные.

Выводы

Предложена модель генерации форм прошедшего времени английских глаголов, основанная на многослойном персептроне, где орфографическая запись конца глагола ассоциируется с аддитивными правилами. В отличие от моделей, основанных на фонетической записи глаголов, данная искусственная нейронная сеть может быть непосредственно использована для автоматической генерации форм известных и новых глаголов. ИНС обучается и тестируется на большом лексиконе глаголов американского варианта английского языка. Приводится анализ этапов обучения и тестирования ИНС, а также предлагаются решения по дальнейшему совершенствованию процесса обучения с целью более эффективного прогнозирования форм глаголов.

Предложенные рекомендации по порядку ввода глаголов и примеров сочетания корней при образовании сложных неправильных глаголов могут быть также использованы при разработке учебников и методических пособий по обучению английскому языку школьников и студентов.

Список литературы

1. Хомский, Н. Три модели описания языка: Пер. с англ. // Кибернетический сборник. — М:ИЛ, 1961. — Вып. 2. — С. 237-266.
2. Хомский, Н. Синтаксические структуры // Новое в лингвистике. — М., 1962. — Вып. II.
3. Фодор, Дж., Пылишин, 3. Коннекционизм и когнитивная структура: критический обзор // Язык и интеллект. — М.: Прогресс, 1996. — С. 230-313.
4. Pinker, S. Words and rules // Lingua. — 1998 — №106, — P. 219-242.
5. Pinker, S, Ullman, M. The past and future of the past tense / Steven Pinker // Trends in Cognitive Science. — 2002. — №6. — P. 456-463.
6. McClelland, J. L, Rumelhart, D.E., Hinton, G.E. The Appeal of Parallel Distributed Processing // Parallel Distributed Processing: Explorations in the Microstructure of Cognition. — Cambridge, MA: MIT Press, 1986. — V. I. — P. 3-44.
7. Rumelhart, D.E., McClelland, J.L. On Learning the Past Tenses of English Verbs // In Parallel Distributed Processing: Explorations in the Microstructure of Cognition. — Cambridge, MA: MIT Press, 1986. — V. 2. — P. 216-271.
8. Pinker, S., Prince, A. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition // Cognition. — 1988. — №28. — P. 73–193.
9. McClelland, J. L., Patterson, K. Rules or Connections in Past-Tense inflections: What does the evidence rule out? // Trends in Cognitive Sciences. — 2002. — №6. — P. 465-472.
10. Plunkett, K.A., Juola, P. Connectionist Model of English Past-Tense and Plural Morphology // Connectionist Psycholinguistics. — Westport, CT; London: Ablex Publishing, 2001. — P. 106-137.
11. Antworth, E.L. Morphological Parsing with a Unification-based Word Grammar // North Texas Natural Language Processing Workshop, May 23, 1994, University of Texas at Arlington. – Режим доступа: http://www.sil.org/pckimmo/ntnlp94.html.
12. Kučera, H., Francis, W.N. Computational analysis of present-day American English. Providence, RI: Brown University Press, 1967.