ВЕСТНИК ПЕРМСКОГО УНИВЕРСИТЕТА

2012 РОССИЙСКАЯ И ЗАРУБЕЖНАЯ ФИЛОЛОГИЯ Вып. 3(19)

УДК 81’33

О КОРПУСНОМ ПОДХОДЕ К ИЗУЧЕНИЮ ОШИБОК ПРИ БИЛИНГВИЗМЕ

Лиана Магдановна Ермакова
аспирант кафедры процессов управления и информационной безопасности
Пермский государственный национальный исследовательский университет
614990, Пермь, ул. Букирева, 15. liana87@mail.ru

Статья посвящена корпусному подходу изучения ошибок, возникающих при билингвизме, а также типологии речевых сбоев. В рамках исследования было реализовано веб-приложение для сбора данных. Программа представляет собой систему аннотирования корпусов текстов. Предусматривается трехуровневая классификация ошибок. Корпус содержит тексты на различных языках (русском, французском, английском, испанском, китайском), написанные как носителями языка, так и искусственными билингвами, и постоянно пополняется. Предлагается метод определения уровня языковой компетенции автора и аутентичности текста (создан он носителем языка или билингвом) на основе машинного обучения.

Ключевые слова: билингвизм; искусственный билингвизм; речевой сбой; ошибка; типология ошибок; корпус текстов; классификация текстов; машинное обучение.

Введение

Бурный интерес ученых к проблемам билингвизма обусловлен рядом социальных, культурных и исторических причин, среди которых следует назвать существование множества билингвальных и полилингвальных государств, массовую миграцию, а также тенденции глобализации. Превращение мира в «глобальную деревню» не только в значительной мере выровняло образ жизни представителей различных этнокультурных сообществ, но и кардинально повлияло на судьбу языков, прежде всего английского. В случае сосуществования и функционирования разных языков в одном государстве проявляется естественный национальный билингвизм, в то время как глобализация стимулирует развитие искусственного индивидуального билингвизма, т.е. учебного двуязычия. Изучение учебного, или искусственного, билингвизма не менее актуально, чем исследование закономерностей развития и функционирования билингвизма естественного [Овчинникова 2011].

Статистические данные свидетельствуют о том, что число билингвов в современном мире неуклонно растет [Терехова 2006]. В связи с этим возникает необходимость разработки и совершенствования стандартных языковых тестов, измеряющих уровень владения языком прежде всего при искусственном билингвизме и диагностирующих степень готовности человека к социализации в языковом коллективе. Появление стандартных языковых тестов, таких как IELTS, TOEFL, DALF и проч., усилило попытки разработки стандартизированных методов обучения иностранным языкам. Несмотря на то что возможность существования подобных методик обосновывается рядом серьезных аргументов, нельзя не принимать во внимание различия между языками. Обучение языкам всегда опиралось на контрастивную лингвистику, обеспечивающую теоретическую основу для разработки дидактического материала, повышающего эффективность обучения и снижающего вероятность ошибок, обусловленных межъязыковой интерференцией. Соответственно, ошибка оставалась основным критерием оценки овладения языком и диагностики слабых звеньев в языковой компетенции билингва.

Ошибка – это одно из важнейших вспомогательных средств овладения языком. Ошибки дают ключ к пониманию механизма овладения языком, поскольку позволяют определить, на каком этапе порождения высказывания происходит сбой, вызван ли сбой интерференцией языков, возможны ли подобные ошибки в речи носителей языка. В случае искусственного билингвизма под ошибкой обычно понимают *«результат неправильной операции выбора языковых средств иностранного языка»* [Будренюк, Григоревский 1978: 30].

Согласно результатам исследований М.В.Русаковой «отклонения в русской [устной] речи являются обыденным явлением, по частотности сопоставимым с таким явлением, как употребление существительных» [Русакова 2007: 60]. Несмотря на то что ошибки и нестандартная речь больше свойственны повседневному общению, чем официальным выступлениям и письменным формам речи, в условиях билингвизма, например в русскоязычном СМИ (особенно Интернете) в Израиле, ошибки встречаются достаточно часто [Ovchinnikova 2011].

Исследования естественного билингвизма и двуязычия в прикладном аспекте зачастую опираются на анализ ошибок, которые допускают дети и взрослые билингвы в различных коммуникативных ситуациях [Cantone 2007; Goldrick 2006; Hartsuiker 2006].

Анализ ошибок широко используется в онтолингвистике, афазиологии и других направлениях психолингвистики. Таким образом, изучение ошибок актуально не только с точки зрения билингвизма, но и с позиций психолингвистики и теории коммуникации в целом. Опыт интерпретации ошибок и механизмов их возникновения позволяет конкретизировать модель порождения и восприятия речи в теоретическом плане и отрабатывать адекватные различным задачам методики диагностики и коррекции речевого и когнитивного развития, формирования коммуникативной компетенции в прикладном аспекте. Современной наукой накоплен богатый опыт изучения и описания ошибок. Между тем на практике часто возникает проблема быстрого определения типичности или специфичности ошибки, диагностики и исправления ошибок автоматически сгенерированных текстов на разных языках, в частности при автоматическом переводе, для чего необходим не столько точечный анализ квалифицированного лингвиста, сколько сопоставление речевой продукции с базой данных ошибок в текстах на разных языках, созданных как носителями языка, так и билингвами. С этой целью мы разработали веб-приложение, позволяющее создавать и размечать корпус ошибок, а также осуществлять поиск ошибок на различных языках. Корпус содержит тексты на различных языках (русском, французском, английском, испанском, китайском и т.д.), написанные как носителями языка, так и изучающими иностранный язык, т.е. искусственными билингвами, и постоянно пополняется. Определяя наиболее проблематичные области и типичные виды ошибок, можно скорректировать обучение иностранному языку. Таким образом, сбор ошибок важен как для конкретной ситуации обучения (для одной группы и одного преподавателя), так и для моделирования процесса овладения языком в принципе. Другой актуальной исследовательской задачей, которую можно решать на основе корпуса, является определение уровня языковой компетенции автора и аутентичности текста: создан он носителем языка или билингвом.

Для создания веб-приложения необходимо было разработать классификацию текстов. В нашем случае речь идет о бинарной классификации, т.е. есть о двух непересекающихся категориях: текстах, написанных носителями языка, и текстах, язык которых не является родным для автора. Задача заключается в отнесении текста к той или иной категории в зависимости от содержащихся в нем ошибок. Обычно классификация документов широко применяется в таких областях, как системы документооборота, фильтрация текстов (например, спама), автоматическое аннотирование и реферирование, машинный перевод, составление интернет-каталогов, ограничение области поиска в поисковых системах, определение кодировки и языка текста, классификация новостей и т.д. [Лифшиц 2005; Sebastiani 2002; Chakrabarti 2003]. Классификация может осуществляться вручную на основе правил или автоматически. Примером ручной классификации может служить библиотека, где книгам вручную присваиваются тематические рубрики. Ручная категоризация неприменима, если необходимо классифицировать большой объем документов за ограниченное время, потому традиционно классификационные задачи решаются методами машинного обучения, такими как многослойный персептрон, машина опорных векторов и байесовские сети. В отличие от перечисленных методов, самоорганизующиеся карты Кохонена являются примером методов кластеризации, т.е. обучения без учителя. Основным отличием категоризации от кластеризации является то, что рубрики заранее задаются пользователем или экспертом, а кластеры формируются автоматически при анализе коллекции [Лифшиц 2005].

Перед тем как охарактеризовать предлагаемый корпус, определим основные теоретические посылки нашего исследования.

Билингвизм

Билингвизм – это явление попеременного использования одним и тем же человеком двух языков [Вайнрайх 1972]. Л.В.Щерба дает следующее определение билингвизма: «Под двуязычием подразумевается способность тех или иных групп населения объясняться на двух языках» [Щерба 1974: 313]. Он подчеркивает социальную направленность языка: «быть двуязычным значит принадлежать одновременно к двум таким различным группировкам». Л.В.Щерба выделяет «чистую» форму билингвизма и «смешанную». В

первом случае социальные группы, лежащие в основе билингвизма, являются взаимоисключающими и языки никогда не употребляются вперемежку. Во втором случае один язык может сменяться другим, а групповые границы стираются. Ученый указывает на множественность промежуточных вариантов между искусственным и естественным двуязычием. Знание языка, по Л.В. Щербе, может быть сознательным и бессознательным, когда мы не думаем о языке, а лишь используем его. Литературный вариант языка всегда отличается от разговорного, поэтому овладение литературной нормой и употребление литературного варианта всегда осознанно [Щерба 1974]. У взрослых преимущественно встречается искусственное усвоение языка, т. е. усвоение в процессе целенаправленного изучения [Терехова 2006]. При изучении иностранного языка человек сравнивает разные формы выражения, таким образом, отделяя мысль от знака [Щерба 1974]. Помимо щербовской классификации билингвизма, разработаны и другие классификации, основанные на соотношении компетенции в контактирующих языках и связанных с языками культур [Вайнрайх 1972]. В зависимости от типа билингвизма мы сталкиваемся с большей или меньшей языковой и культурной интерференцией. Уриэль Вайнрайх выделяет три типа взаимодействия языков:

1. Языковой сдвиг. Первый язык может быть заменен вторым языком.
2. Переключение. При переключении оба языка используются по очереди.
3. Слияние языков в единую языковую систему [Вайнрайх 1972].

В последние годы появились работы, описывающие различные промежуточные типы:

- литературный (писательский) билингвизм;
- переводческий билингвизм, семейный (унаследованный билингвизм) [Нестерова, Овчинникова 2012].

Характерной особенностью билингвизма, в особенности искусственного двуязычия, является интерференция. Интерференция – это взаимодействие языковых систем в условиях дву- или многоязычия, обусловленное их структурными расхождениями и проявляющееся в отклонении от кодифицированных норм речи контактирующих языков. Интерференция отражает как взаимодействие языков, так и некоторую промежуточную стадию формирования языковой компетенции на неродном языке. В настоящее время преобладает гипотеза «промежуточного языка» («интерязыка»), согласно которой при формировании билингвизма у человека в каждый момент времени есть система представлений об изучаемом языке (Я2). Эта система может постепенно приближаться к идеальной языковой системе Я2. Очевидно, что наиболее эффективным способом подтверждения этой гипотезы или выдвижения новой будет система, позволяющая регистрировать ошибки билингва на разных периодах обучения: ошибка рассматривается как *источник информации о процессе овладения Я2* [Залевская 2007]. При этом не стоит забывать и о прикладном аспекте изучения билингвизма: определяя наиболее проблематичные области и типичные виды ошибок, мы можем скорректировать обучение иностранному языку. Таким образом, сбор ошибок важен как в конкретной ситуации обучения (для одной группы и одного преподавателя), так и для моделирования процесса овладения языком в принципе.

Однако система регистрации ошибок будет бесполезной без классификации и типологии ошибок и их возможных причин. Классификация, лишенная причин возникновения, иррелевантна для методики обучения иностранному языку, поскольку ошибки интерференции не отграничиваются от ошибок, возникающих по другим причинам, а «незнание причины возникновения ошибок лишает педагога возможности организовать целенаправленную работу по их устранению и предупреждению» [Карлинский 1989].

Понятие речевого сбоя и типология ошибок

Первые списки ошибок были опубликованы в конце XIX в. Мерингерем и Майером [Русакова 2007]. Современную эпоху в области изучения речевых ошибок открыли работы В. Фромкин [Fromkin 1973]. Именно В. Фромкин предложила один из первых корпусов ошибок, что позволило определить вариативность ошибок на различных уровнях языка.

Типология ошибок в современной лингвистике представлена множеством вариантов. Начиная с С.Кордера [Corder 1967, 1974, 1981] принято последовательно разграничивать оговорки/описки, связанные с усталостью или невнимательностью, и собственно ошибки (mistakes и errors). Г.М.Будренюк и В.М.Григоревский также выделяют оговорки и логические ошибки, связанные с выбором средств [Будренюк, Григоревский 1978]. Последние и являются настоящим предметом исследования. Рассмотрим одну из типологий – классификацию ошибок в речи, предложенную Стембергером [Stemberger 1985]. Он разделяет ошибки

- по обусловливающим их процессам (замены, смешения, добавления, опущения);
- по тому, какой из языковых уровней оказался нарушенным (синтаксис, морфология,

морфолого-синтаксический уровень, фонетика, лексика);

- по тому, связаны ли они с контекстом (антиципация, персеверация, транспозиция, сдвиг) или характеризуются как внеконтекстные, а также по отношению между источником и местом ошибки (внутри одной единицы или между единицами).

Такая классификация уже включает в себя причину ошибок (предвосхищение, замена элементов и т.д.), мы же предлагаем указать отдельно на тип ошибки (например, морфологическая – неправильно выбран род – женский вместо мужского) и на ее возможную причину (например, интерференция с родным языком билингва).

Заметим, что данная классификация создана для ошибок в устной коммуникации, но во многом подходит и для ошибок, допущенных в письменных текстах. Поскольку в предлагаемой нами системе возможно использование как письменных текстов (сочинения учащихся), так и устных (расшифровка диалогов и монологов), то ее вполне можно принять за основу. Далее каждый пользователь или преподаватель сможет дополнять систему классификации ошибок.

По частоте встречаемости в речи разных носителей языка ошибки могут быть типичными или окказиональными [Балясникова 2007].

Ошибки мышления, возникающие при восприятии речи, могут быть следствием ложного прогнозирования, связанного с возникновением ложных ассоциаций с прежним опытом и апперцепции. Невнимательность, интерференция и сверхгенерализация встречаются как при восприятии речи, так и при ее порождении, например, буквосочетание *-ea-* может быть прочитано изучающими английский язык как [i:], в том числе в словах *dead, deaf, death, break*.

В рамках данного исследования для нас больший интерес будут представлять ошибки, возникающие при производстве речи.

Причины возникновения речевых ошибок можно классифицировать следующим образом [там же]:

- лингвистические (интерференция, сверхгенерализация);
- экстралингвистические (неправильные методические действия, условия обучения);
- комплекс первых двух типов.

Речевые ошибки, возникающие при изучении иностранного языка, часто возникают по причине межъязыковой интерференции. Обусловленные интерференцией ошибки можно спрогнозировать [там же]. Интерференция часто связана с недостаточной автоматизированностью употребления языковых единиц, вследствие чего менее усвоенные элементы заменяются более усвоенными, но принадлежащими другой языковой системе [Будренюк, Григоревский 1978]. Например, отсутствие артиклей в русском языке приводит к тому, что многие носители русского языка, изучающие английский или любой другой язык, в котором артикль выражается регулярно, в большинстве случаев не знают, в каких случаях надо употреблять определенный, неопределенный или нулевой артикль. В этом случае можно говорить о грамматической интерференции. По мнению У.Вайнрайха, «грамматическая интерференция возникает тогда, когда правила расстановки, согласования, выбора или обязательного изменения грамматических единиц, входящих в систему языка S, применяются к примерно таким же цепочкам элементов языка C, что ведет к нарушению норм языка C, либо тогда, когда правила, обязательные с точки зрения грамматики языка C, не срабатывают ввиду их отсутствия в грамматике языка S» [Вайнрайх 1972: 35]. Грамматическая интерференция может также проявляться в неправильном порядке слов. Например, носители русского языка (или другого языка с относительно свободным порядком слов) могут допустить следующую ошибку: «*he comes tomorrow home*». Другим примером грамматической интерференции может быть ошибочное употребление формы прилагательного при согласовании в роде с существительным («советский общество») носителем китайского языка, изучающим русский язык. Ошибки также могут быть вызваны несовпадением грамматических категорий, например, носители русского языка при изучении французского могут ошибочно определить род таких слов греческого происхождения, как система, которые относятся к женскому роду в русском языке и к мужскому во французском («*le système*»). Аналогичным образом грамматическая система русского языка может влиять на правильное употребление глагольных форм в английском языке: «*if I shall do it*». Другой не менее важной причиной ошибок может быть сверхгенерализация, например, проявляющаяся при выборе формы нерегулярных глаголов («*taked*» вместо «*took*»), образования формы множественного числа («*childrens*», «*corpuses*» вместо «*corpora*», «*phenomenons*» или «*phenomenas*»вместо «*phenomena*»), степени сравнения прилагательных («*more good*») и т.д.

С другой стороны, носители языка также делают ошибки в родном языке. К наиболее частым ошибкам носителей языка можно отнести речевые ошибки, такие как непонимание значения слова или многословие. Многословие включает в себя плеоназм (употребление в речи близких по смыслу, логически излишних слов)

(например, «*бесплатный подарок*», «*разных точках зрения ученых, у которых различаются мнения*»), тавтология («*писатель писал*»), расщепление сказуемого («*машина пошла на разворот*»), использование лишних слов и слов-паразитов («*как бы*», «*типа*»). Плеоназм и тавтология отличаются от использования лишних слов и слов-паразитов тем, что в первом случае лексическое значения уже выражено с помощью других слов, а во втором эти слова просто не нужны в данном контексте.

Интерференция языков может идти в обоих направлениях: Я1 влияет на Я2 и наоборот. Часто влияние второго языка проявляется в трансформации фразеологизмов и устойчивых сочетаний. Под влиянием английского языка и иврита выражение «*требует времени*» преобразуется в «*берет время*» по аналогии с «*takes time*» в речи русских израильтян [Ovchinnikova 2011]. Замечено, что билингвы часто искажают фразеологизмы не только под влиянием интерференции; трансформация фразеологизмов обычно сводится к пропуску элементов («*цыплят считают*» вместо «*цыплят по осени считают*») и замене редкого слова словом более широкой семантики («убирать взор» вместо «*отводить взор*») [там же]. Это объясняется делексикализацией фразеологизмов в условиях билингвизма. Многие идиомы теряют свою целостность: они реконструируются, а не хранятся в памяти как единое целое [Philip 2007].

Лексические ошибки могут быть вызваны следующими причинами:

- неправильным отождествлением слов из-за несоблюдения звуковых и интонационных различий языков (англ. «*mourning*» (горевать) вместо англ. «*morning*» (утро), англ. «*bier*» (гроб) вместо англ. «*beer*» (пиво));
- наличием коррелирующих когнатов в разных языках или интернациональными словами («*accurate*» вместо «*neat*», «*precise*»);
- прямым лексическим переносом;
- наличием ложных когнатов (англ. «*bless*» (благословлять) и фр. «*blesser*» (ранить), англ. «*pain*» (боль) и фр. «*pain*»(хлеб));
- многозначностью (использованием одной лексемы одного языка, которой соответствует ряд лексем другого языка) [Балясникова 2007].

Большая часть ошибок билингвов в сочетаемости единиц связана с калькированием: по аналогии с «*Nowadays with the rapid developments*» образуется «*в наши дни со стремительным развитием*», «*we are regulated by EU*» – «*мы регулированные Евросоюзом*», «*делать диссертацию*», в речи русских израильтян представляет собой кальку с иврита [Ovchinnikova 2011]. Иногда смыслоразличительную функцию выполняют грамматические показатели, например артикль (фр. «*bierre*» в женском переводится на русский язык «*пиво*», а в мужском – «*гроб*»), что также приводит к ошибкам. Однако неправильный выбор лексической единицы (неразличение паронимов) может характеризовать как речь билингвов, так и носителей языка («*заглавный герой*» вместо «*главный герой*») [ibid.].

Ошибки на морфологическом уровне могут быть вызваны интеграцией лексического и грамматического значений слова (например, существительные греческого происхождения во многих языках принадлежат мужскому роду, но имеют форму, схожую с формами женского рода: *исп. el systema*). Грамматические ошибки могут быть вызваны неадекватным анализом (управление глагола, подчинение не реализованному впоследствии глаголу и т.д.) («*Я сейчас в Китай*», «*Я люблю рыба*», «*Это было совершенно не свойственно для взрослых*»), интроспекцией («*он не зачислился в списки*») и т.п. [Русакова 2007]. Подвержен ошибкам, возникающим под влиянием интерференции языков в сознании билингва, синтаксис. И.Г. Овчинникова выделяет 6 типов синтаксических ошибок, связанных с явлением билингвизма:

1. Порядок слов («*at the teacher's table should be computer*»).
2. Усложнение синтаксиса («*работаю тем, что преподаю дидактику*»).
3. Пересечение зависимостей.
4. Ошибки в сложном предложении («*чем больше языков знает человек, то она культурная и развита*»).
5. Нарушения в выборе временной формы глагола («*проявляет чересчур большую любознательность, за что чуть не поплатился жизнью*»).
6. Пропуск глагола («*he good*») [Ovchinnikova 2011].

Девиантные формы могут быть спровоцированы влиянием предыдущего контекста, содержащего другие алломорфы с тем же корнем [Русакова 2007] («*Связала варежки. Носки связу на следующей неделе*»), или влиянием предыдущих грамматических форм [Ovchinnikova 2011] («*восприятие реальности Маши*» вместо «*Машей*») в речи как билингвов, так и носителей языка.

Таким образом, исследователи рассматривают как оговорки/описки, связанные с усталостью или невнимательностью, и собственно ошибки. Сбои проявляются на всех языковых уровнях в речи как носителей языка, так и билингвов. Оговорки могут быть спровоцированы контекстом (влияние прайма, подчинение нереализованному

глаголу и т.д.). В условиях билингвизма ошибки чаще всего вызваны интерференцией языков, т.е. влиянием Я1 на Я2 и наоборот.

Современные методы исследования ошибок

С конца XX в. исследование ошибок смещается от анализа отдельных документов к текстовым корпусам, от общих задач к конкретным. Большинство существующих работ посвящено анализу отдельных ошибок, однако единственный сбой не всегда позволяет определить истинные причины его появления. Перед современной лингвистикой стоит задача выявить скрытые механизмы реализации речи, что возможно лишь при наличии репрезентативного корпуса речевых отклонений [Русакова 2007].

Корпусный подход позволяет анализировать ошибки, в том числе и с точки зрения их дистрибуции в различных коммуникативных ситуациях. Кроме того, корпусный метод дает возможность выявить как типичные ошибки, так и окказиональные. Анализ аномальных ошибок позволяет выявить скрытые тренды в развитии языка.

В последние десятилетия активно разрабатываются и пополняются корпусы ошибок. Общий список корпусов доступен на http://www.ldc.upenn.edu/Catalog/index.jsp. Наиболее известной является база данных, составленная Викторией Фромкин [Fromkin 2010]. База позволяет работать с английским, французским, немецким и итальянским языками. Наиболее продуктивным корпусный подход оказывается для изучения отступлений от нормативных и типичных коллокаций в текстах, созданных в условиях учебного двуязычия [Philip 2007].

В рамках исследований католического университета Лувена был разработан ряд корпусов, содержащих тексты, созданные студентами, изучающими английский язык:

- International Corpus of Learner English ICLE [Granger 2011] состоит из эссе, написанных носителями различных языков с высоким уровнем владения английским.
- French Interlanguage Database FRIDA [Granger 2010] содержит тексты изучающих французский язык. Корпус разделен на три части: (1) тексты, написанные носителями английского языка, (2) тексты, написанные носителями немецкого языка, и (3) все остальные тексты.
- Louvain International Database of Spoken English Interlanguage LINDSEI [Granger 2012] представляет собой корпус разговорной речи, произведенной носителями различных языков с высоким уровнем английского.
- Longitudinal Database of Learner English LONGDALE [Meunier 2010] является диахроническим корпусом, состоящим из текстов, произведенных студентами Лувена на протяжении изучения английского языка в течение трех и более лет.
- Varieties of English for Specific Purposes dAtabase VESPA [Paquot 2011] состоит из текстов различной тематики и жанровой направленности.

Все корпусы, разработанные католическим университетом Лувена, являются закрытыми коммерческими проектами, что не позволяет свободно использовать их материалы исследователями ошибок.

Еще одним примером корпуса ошибок, возникающих при изучении английского языка, является совместный проект Автономного университета Мадрида и Политехнического университета Валенсии TREACLE (Teaching Resource Extraction from an Annotated Corpus of Learner English). Корпус включает в себя тексты, написанные испанцами при изучении английского языка.

Информация об ошибках и отклонениях в русской речи представлена в разметке русского речевого корпуса [Русакова 2007], также не выставленного в открытый доступ.

Таким образом, корпусы ошибок, возникающих при билингвизме, необходимы для выявления систематических сбоев при изучении иностранных языков. Существующие корпусы являются закрытыми и недоступны для исследований. Многие корпусы содержат тексты на английском языке. В то же время не существует корпусов, содержащих билингвальные ошибки, возникающие в русском языке или под его влиянием. В связи с этим целесообразна разработка простой пополняемой базы, содержащей размеченные тексты, соответствующие редким парам языков.

Описание разрабатываемого корпуса ошибок

В рамках исследования было реализовано вебприложение для сбора данных. Программа представляет собой систему аннотирования корпусов текстов.

Система предусматривает трехуровневую классификацию ошибок. На первом уровне, обозначенном как тип ошибки, находятся следующие категории:

- лексика;
- орфография;
- морфология;
- синтаксис;
- пунктуация;
- стилистика;
- речевые ошибки.

На втором уровне, которому в терминах программы соответствует подтип, присутствуют бо-

лее частные категории ошибок, например, ошибки в согласовании, управлении, образовании форм различных частей речи и т.д. На третьем уровне (специфический тип ошибки) находятся наиболее конкретные типы ошибок (например, отсутствие артикля, управление некоторого предлога, согласование в роде, числе и падеже, неверное образование сравнительной степени прилагательного и т.п.).

В отличие от большинства систем аннотирования текстов, разработанное нами приложение не ограничивает возможности аннотаторов в выборе типа ошибок. Классификация ошибок создается во время процесса разметки текстов. Это позволяет учитывать только реально присутствующие в текстах типы ошибок, которые влияют на статистику. Если первый уровень классификации является практически универсальным для всех языков, то второй и еще в большей степени третий уровни зависят от национального языка. Таким образом, возможность изменения классификации обеспечивает переносимость приложения на другие языки. Кроме того, любая классификация не полна или не точна, а разработанное приложение позволяет расширять типологию по мере необходимости.

В случае использования программы в качестве образовательного портала возможность добавления собственной типологии ошибок позволяет преподавателям иностранных языков адаптировать приложения под конкретные задачи обучения.

При работе с корпусом можно варьировать параметры поиска. Любую ошибку в речи билингва можно сравнить с уже описанными и представленными в корпусе и тем самым выявить ее типичность или, напротив, уникальность.

Любой пользователь имеет возможность скопировать файл со статистическими данными о количестве и типе ошибок в каждом тексте, содержащий также информацию о языке, на котором написан текст, и родном языке автора текста. Статистика представляет собой файл в формате CSV (Comma Separated Value), который является универсальным форматом представления данных и может быть использован в большинстве существующих классификаторов, например WEKA или STATISTICA. Формат удобен для просмотра, т.к. открывается стандартными программами, а также MS EXCEL.

Классификация

Постановка задачи классификации

При классификации документ относится к одному из заранее заданных классов, в отличие от кластеризации, где кластеры формируются автоматически при анализе коллекции [Лифшиц 2005]. Пусть дано конечное множество категорий $C = \{c_1, c_2, \dots, c_{|C|}\}$ и конечное множество документов $D = \{d_1, d_2, \dots, d_{|D|}\}$. Целевая функция $A: D \times C \to \{0,1\}$, которая для каждой пары <документ, категория> определяет, соответствуют ли они друг другу, не известна. Необходимо найти классификатор a, т.е. функцию, максимально близкую к функции A [Лифшиц 2005].

В нашем случае речь идет о бинарной классификации, т.е. есть о двух непересекающихся категориях: текстах, написанных носителями языка, и текстах, язык которых не является родным для автора. Задача заключается в отнесении текста в той или иной категории в зависимости от содержащихся в нем ошибок.

При классификации мы использовали свободное программное обеспечение для анализа данных, написанное на Java, Weka (Waikato Environment for Knowledge Analysis). Для оценки обобщающей способности выбранных моделей применялся механизм кросс-валидации (скользящего контроля). Этот метод применим для алгоритмов, обучаемых по прецедентам. В Weka существует три типа кросс-проверки [Remco 2008]:

1. Контроль по отдельным объектам (leave one out). Для формирования обучающей выборки из исходных данных удаляется *i*-й элемент, который представляет собой тестовую выборку.
2. Контроль по K блокам (K-fold). Исходные данные *D* разбиваются на *m* примерно равных частей $D_1, \dots, D_m$. Обучающая выборка формируется путем удаления блока D_i, который становится контрольной выборкой.
3. Кумулятивный контроль (cumulative). При кумулятивном подходе элементы последовательно добавляются в обучающую выборку.

Мы выбрали контроль по 10 блокам.

Байесовская сеть

Вероятностная постановка задачи классификации заключается в следующем: по заданной обучающей выборке $X^l = (d_i, c_i)_{i=1}^{|X^l|}$ необходимо построить классификатор с минимальной вероятностью ошибки [Воронцов 2011]. В практических задачах минимизируют средний риск, т.е. математическое ожидание потерь для классификатора $a: D \to C$, разбивающего D на непересекающиеся области $A_c = \{d \in D | a(d) = c\}, c \in C$:

$$a(d) = \arg\min R(a) = \arg\min \sum_{c \in C} \sum_{s \in C} \lambda_{cs} P(A_s, c) = \arg\min_{s \in C} \sum_{s \in C} \lambda_{cs} P(c) p(d|c)$$
$$= \arg\max_{c \in C} \lambda_c P(c) p(d|c) \text{ при } \lambda_{cc} = 0, \lambda_{cs} = \lambda_c$$

где $P(A_s, c)$ – вероятность ошибки, т.е. объект d класса c попадает в $A_s, s \neq c$; $\lambda_{cs} \geq 0$ – потеря от ошибки; $P(c)$ – априорная вероятность класса c, а $p(d|c)$ – функция правдоподобия класса c [Воронцов 2011]. При этом априорные вероятности оцениваются по частоте встречаемости в обучающей выборке [Воронцов 2011]:

$$\hat{P}(c) = \frac{l_c}{l}.$$

Байесовская сеть B над объектами $U = \{x_1, x_2, \ldots, x_n\}, n \geq 1$ представляет собой структуру B_S в виде ориентированного ациклического графа и множество вероятностных таблиц $B_P = \{p(u|pa(u)|u \in U\}$, где $pa(u)$ – множество родителей вершины u в B_S [Remco 2008].

При классификации мы ограничили число родительских вершин до 1. Для решения оптимизационной задачи использовали алгоритм K2 с байесовской метрикой оценки качества сети. Условные вероятности вычислялись как

$$P(x_i = k | pa(x_i) = j) = \frac{N_{ijk} + N'_{ijk}}{N_{ij} + N'_{ij}},$$

где N_{ijk} – количество раз, когда $pa(x_i) = j$, а $x_i = k$; N_{ij} – количество раз, когда $pa(x_i) = j$, N'_{ij} и N'_{ijk} – параметры. Мы использовали значения параметров по умолчанию $N'_{ij} = N'_{ijk} = 0.5$.

Байесовская сеть правильно классифицировала 92,27% объектов.

При наивном подходе предполагается независимость признаков. В этом случае функции правдоподобия классов представимы в виде произведения одномерных плотностей признаков:

$$p(d|c) = \sum_{i=1}^{n} p(\xi_i)$$
$$d = (\xi_1, \ldots, \xi_n).$$

Наивный байесовский классификатор показал самые низкие результаты – 55,55% правильно классифицированных текстов.

Машина опорных векторов

Метод опорных векторов (SVM – support vector machine) – это набор схожих алгоритмов на основе обучения с учителем, применяющийся для анализа данных и распознавания образов в задачах классификации и регрессионном анализе. SVM является линейным классификатором. На основе обучающей выборки алгоритм помогает предсказать, в какую из двух категорий попадает элемент, подлежащий классификации. Основная идея – построение гиперплоскости или набора гиперплоскостей в пространстве более высокой размерности и максимизация расстояния между построенной гиперплоскостью и классами обучающей выборки. Переход в спрямляющее пространство более высокой размерности требуется, т.к. в общем случае выборка может не быть линейно разделимой. Функция перехода называется ядром. В настоящее время не существует общего алгоритма подбора ядер [Воронцов 2011].

В качестве ядра мы использовали радиальные базисные функции. Результат SVM составил 86,47%.

Нейронные сети

Искусственная нейронная сеть – это математическая модель, представляющая собой систему искусственных нейронов. Нейроны функционируют локально. Они соединены при помощи синапсов (однонаправленных связей). Обучение нейронной сети – это многопараметрическая задача нелинейной оптимизации. В большинстве случаев нейронные сети являются адаптивными системами, меняющими свою структуру в зависимости от внутренней и внешней информации, поступающей во время процесса обучения [Горбань и др. 1998: 296].

Многослойный персептрон

Многослойный персептрон является искусственной нейронной сетью прямого распространения, т.е. все связи направлены строго от нейронов входного слоя к нейронам выходного слоя, при этом каждый нейрон i-го слоя связан со всеми нейронами i+1-го слоя. Все нейроны, кроме нейронов входного слоя, имеют нелинейную функцию активации (обычно сигмоидальную). Для обучения используется алгоритм обратного распространения ошибки [Rumelhart, Hinton, Williams 1986; Rosenblatt 1961].

Несмотря на то что, в отличие от машины опорных векторов, нейронные сети в общем случае не гарантируют максимальный зазор между разделяющей гиперплоскостью и классами, персептрон показал самые высокие результаты классификации – 97,58% правильно классифицированных текстов. Вероятно, это связано с тем, выборка не является линейно разделимой и необходимо подобрать оптимальное спрямляющее пространство. К сожалению, не существует универсального метода подбора базисных функций, и выбранные радиальные базисные функции показали результат хуже, чем многослойный персептрон.

Самоорганизующиеся карты Кохонена

Самоорганизующиеся карты Кохонена (self orginizing maps, SOM) представляют собой соревновательную нейронную сеть с обучением без учителя. В отличие от других нейронных сетей, SOM использует понятие расстояния для сохранения топологической структуры исходного пространства. Каждый нейрон имеет координаты на карте и вектор весовых коэффициентов. В качестве топологии обычно используется регулярная прямоугольная или шестиугольная решетка. При обучении на вход сети подается векторное представление объекта и вычисляется расстояние до каждого вектора весовых коэффициентов, после чего вычисляются новые значения весов по формуле

$$Wv(t+1) = Wv(t) + \Theta(v,t)\,\alpha(t)(D(t) - Wv(t)),$$

где $\alpha(t)$ – монотонно убывающая функция, задающая скорость обучения; $D(t)$ – вектор входных параметров; $\Theta(v,t)$ – функция соседства. Обычно в качестве функции соседства применяют гауссовскую функцию:

$$\Theta(v,t) = e^{\frac{\|r_v - r_i\|}{2\sigma^2(t)}},$$

где r_v, r_i – координаты узлов на карте; $\sigma(t)$ – монотонно убывающая функция, определяющая количество соседей. При классификации выбирается нейрон с минимальным расстоянием до входного вектора [Kohonen 2001].

Карты Кохонена также дали очень хорошие результаты – 95,65% правильно определенных объектов.

Заключение

Актуальность изучения ошибок при билингвизме и учебном двуязычии обусловлена двумя основными факторами. Первым фактором является необходимость оптимизации коммуникации и обучения языкам в условиях межкультурной коммуникации. В качестве второго фактора стоит назвать целесообразность конкретизации описания механизмов речи.

Исследователи разграничивают оговорки/описки, связанные с усталостью или невнимательностью, и собственно ошибки. Сбои проявляются на всех языковых уровнях как в речи носителей языка, так и билингвов. Оговорки могут быть спровоцированы контекстом (влияние прайма, подчинение нереализованному глаголу и т.д.). В целом для билингвов характерны межъязыковые ошибки, обусловленные интерференцией языков, причем при учебном билингвизме таких ошибок больше.

На наш взгляд, характеристики речи при разных типах билингвизма различаются не столько уникальными дифференциальными признаками, сколько вариативностью распределения и частотой встречаемости общих признаков. Ошибки наблюдаются у всех билингвов, но частотность ошибок разных типов не совпадает.

Корпусы ошибок, возникающих при билингвизме, необходимы для выявления систематических сбоев при изучении иностранных языков. В свою очередь, анализ аномальных ошибок позволяет выявить скрытые тренды в развитии языка. Существующие корпусы являются закрытыми и недоступны для исследований. Практически все корпусы ориентированы на анализ ошибок, возникающих при изучении английского языка. В связи с этим целесообразна разработка простой пополняемой базы, содержащей размеченные тексты, соответствующие редким парам языков.

Ошибки и оговорки в речи билингвов отличаются по типу и частотности от тех, что допускают монолингвы. Между тем в последние годы положительный перенос когнитивных навыков с родного языка на второй подвергается сомнению: такой перенос очевиден не для всех языков. Обнаружены различия в переносе навыков, приобретенных на родном языке, на обработку информации, поступающей на втором языке; когнитивные навыки переносятся легко, в то время как базовые знания и привычки в межличностной коммуникации часто не подлежат переносу и формируются на основе коммуникативного опыта в новом языковом сообществе [Cummins 1991].

В связи с этим перед нами стоял вопрос, можно ли по имеющимся в тексте ошибкам определить, написан данный текст искусственным билингвом или носителем языка. Для ответа на этот вопрос мы применили методы машинного обучения на корпусе аутентичных текстов и текстов, созданных в условиях искусственного билингвизма. Персептрон показал самые высокие результаты классификации – 97,58% правильно классифицированных текстов. Наивный байесовский классификатор показал самые низкие результаты – 55,55% правильно классифицированных текстов. Результат SVM составил 86,47%, карты – Кохонена 95,65%, а байесовская сеть правильно классифицировала 92,27% объектов. Столь высокие результаты в некоторой степени объясняются особенностью коллекции: большинство текстов, написанных искусственными билингвами, содержит значительное количество ошибок, в то время как сбои, встречающиеся в текстах, написанных носителями языка, относительно редки.

Список литературы

Балясникова Н.С. Природа и типология ошибок при изучении английского языка как второго иностранного при первом испанском // Изв. Рос.

гос. пед. ун-та им. А.И.Герцена. 2007. Т.24, №6. С.88–92.

Будренюк Г.М., Григоревский В.М. Языковая интерференция и методы ее выявления. Кишинев: Штиинца, 1978. 126 с.

Вайнрайх У. Одноязычие и многоязычие // Новое в лингвистике. 1972. №6. Языковые контакты. С.25–60.

Воронцов К.В. Статистичесие (байесовские) методы классификации. URL: http://www.machinelearning.ru/wiki/images/9/98/Voron-ML-Bayes-slides.pdf (дата обращения: 12.03.2011).

Горбань А.Н. и др. Нейросетевые информационные модели сложных инженерных систем // Нейроинформатика. Новосибирск: Наука. Сиб. предпр. РАН, 1998. С. 296.

Залевская А.А. Введение в психолингвистику. Изд. 2-е, доп. М: РГГУ, 2007. 558 с.

Карлинский А.Е. Экспериментальное изучение лексической интерференции в прикладных целях. Текст // Сравнительно-сопоставительное изучение языков и интерференция: сб. науч. тр. Алма-Ата, 1989. С.51–60.

Лифшиц Ю. Классификация текстов. 2005. URL: http://yury.name/internet/ (дата обращения: 10.10.2011).

Нестерова Н.М., Овчинникова И.Г. Исследование переводческого билингвизма на материале корпуса ошибок // Жизнь языка в культуре и социуме 3: материалы конф. Москва, 20–21 апреля 2012 г. / ред. кол.: Е.Ф.Тарасов (отв. ред.), Н.Ф.Уфимцева, В.П.Синячкин. 2012. С.327–330.

Овчинникова И.Г. К проблеме специфики ошибок в случае естественного билингвизма: сопоставление речи русско-иврит билингвов с материалами национального корпуса русского языка // Проблемы социо- и психолингвистики: сб. ст. / отв. ред. Е.В.Ерофеева; Перм. гос. нац. исслед. ун-т. Пермь, 2011. Вып.15: Пермская социопсихолингвистическая школа: идеи трех поколений: к 70-летию Аллы Соломоновны Штерн. С.168–182.

Русакова М.В. Сбои при порождении словоформы в устной речи как результат спонтанного взаимодействия стратегий и механизмов // Материалы XXXVI Междунар. филол. конф., 12–17 марта 2007 г., Санкт-Петербург. Вып.20. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 2007. С.59–71.

Терехова А.В. Некоторые проблемы раннего билингвизма в искусственно созданной иноязычной среде // Иностранный язык как предмет преподавания и исследования: сб. метод. и науч. исслед., 2006. С.75–79.

Щерба Л.В. К вопросу о двуязычии // Языковая система и речевая деятельность. Л., 1974. С.313–318.

Cantone K.F. Code-Switching in Bilingual Children. Springer, 2007. 296 p.

Chakrabarti S. Mining the Web: Discovering Knowledge from Hypertext Data. Morgan-Kaufmann Publishers, 2003. 352 p.

Corder S.P. The significance of learner`s errors // International Review of Applied Linguistics. 1967. Vol.5. P.161–170.

Corder S.P. Error analysis // Techniques in Applied linguistics, The Edinborough Course in Applied Linguistics. Oxford: Oxford University Press, 1974. P.122–154.

Corder S.P. Error analysis in interlanguage. Oxford: Oxford University Press, 1981. 125 p.

Cummins J. Language Development and Academic Learning // Language, Culture and Cognition Clevedon: Multilingual Matters. 1991. P.161–175.

Fromkin V. Speech errors as linguistic evidence. The Hague: Mouton, 1973. 269 p.

Fromkin V. Fromkins Speech Error Database. 2010. URL: http://www.mpi.nl/cgi-bin/sedb/sperco_form4.p (дата обращения: 15.09.2011).

Goldrick M. Limited interaction in speech production: Chronometric, speech error, and neuropsychological evidence // Language and Cognitive Processes. 2006. №21. P.817–855.

Granger S. French Interlanguage Database. 2012. URL: http://www.uclouvain.be/en-cecl-frida.html (дата обращения: 20.04.2012).

Granger S. International Corpus of Learner English. 2012. URL: http://www.uclouvain.be/en-cecl-icle.html (дата обращения: 20.04.2012).

Granger S. Louvain International Database of Spoken English Interlanguage. 2012. URL: http://www.uclouvain.be/en-cecl-lindsei.html (дата обращения: 20.04.2012).

Hartsuiker R.J. Are speech error patterns affected by a monitoring bias? // Language and Cognitive Processes. 2006. Vol.21, No 7–8. P.856–891.

Kohonen T. Self-Organizing Maps. N.Y., 2001. 501 p.

Meunier F. Longitudinal Database of Learner English. 2010. URL: http://www.uclouvain.be/en-cecl-longdale.html (дата обращения: 20.04.2012).

Ovchinnikova I. Russian Language in Israel: trends for development // Israel Studies in Language and Society. 2011. No 1. P.15–23.

Paquot M. Varieties of English for Specific Purposes dAtabase (VESPA) learner corpus. 2011. URL: http://www.uclouvain.be/en-cecl-vespa.html (дата обращения: 20.04.2012).

Philip G. Decomposition and delexicalisation in learners' collocational (mis)behaviour // Proceedings of Corpus Linguistics. 2007. P.1–11.

Remco R.B. Bayesian Network Classifiers in Weka for Version 3-5-7. 2008. URL: www.cs.waikato.ac.nz/~remco/weka.bn.pdf (дата обращения: 14.10.2011).

Rosenblatt F. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961. 616 p.

Rumelhart D.E., Hinton G.E., Williams R.J. Learning Internal Representations by Error Propagation // Parallel distributed processing: Explorations in the microstructure of cognition. 1986. Vol.1. P.318–362.

Sebastiani F. Machine learning in automated text categorization // Journal ACM Computing Surveys (CSUR). 2002. Vol.34, No 1. P.1–47.

Stemberger J.P. The lexicon in a model of language production. N.Y.: Garland Publishing, 1985. 215 p.

CORPUS STUDY OF BILINGUAL ERRORS

Liana M. Ermakova
Post-graduate Student of Management Process and Information Security Department
Perm State National Research University

The article is devoted to the corpus study of bilingual errors and to the typology of speech errors. Within the framework of the research we used the web corpus manager that is the system for annotating the corpus of texts. The system provides a three-level error classification. The corpus includes texts in various languages (Russian, French, English, Spanish, Chinese, etc.) written by native speakers as well as by intentional bilinguals. We propose a method of identification of the author's language acquisition competence and authenticity of the text (written by a native speaker or a bilingual) based on machine learning.

Key words: bilingualism; intentional bilingualism; speech errors; errors; error typology; text corpus; text classification; machine learning.