Санкт-Петербургский государственный университет

МЕНЬШИКОВА Алла Павловна

Выпускная квалификационная работа

Предсказание интонационного оформления высказывания для синтеза речи по тексту

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа ВМ.5715. «Общая и прикладная фонетика (General and Applied Phonetics)»
Профиль «Речевые технологии»

Научный руководитель:
доцент, Кафедра фонетики и методики преподавания иностранных языков,
Кочаров Даниил Александрович
Рецензент:
ведущий руководитель, Общество с ограниченной ответственностью «Центр речевых технологий»,
Таланов Андрей Олегович

Санкт-Петербург
2021

Оглавление

Введение

Целью исследования являлась разработка и тестирование метода предсказания интонационного оформления высказывания на русском языке, а именно, таких его аспектов, как расположение интонационного центра внутри синтагмы и ее интонационная модель. Для достижения поставленной цели были решены следующие задачи:

1. Формализация лингвистических признаков для предсказания интонационного оформления, а также отбор наиболее эффективных из них; выбор параметров для нейросетевой модели предсказания.
2. Обучение и тестирование модели предсказания на ряде выборок (дикторозависимых и дикторонезависимых, а также содержащих тексты разного жанра).
3. Лингвистический анализ ошибок модели.

Актуальность исследования обоснована применимостью и полезностью предсказания интонационного оформления для комплексных (End-to-End) систем синтеза речи по тексту, основанных на нейросетевых подходах, которым в последние годы были посвящены многочисленные исследования. Ряд предложенных архитектур и систем, таких как Tacotron2 (Wang et al., 2017), DC-TTS (Tachibana et al., 2018), FastSpeech2 (Ren et al., 2020) и др. (Shen et al., 2918), позволил достичь значительных успехов в повышении качества звучания синтезированной речи, которая в некоторых ситуациях по своей натуральности стала приближаться к человеческой. Современные системы синтеза, зачастую не требующие никаких обучающих данных помимо записей речи и их орфографической расшифровки, предоставляют теоретическую возможность снизить количество требуемой ручной разметки материала до минимума, а также отказаться от лингвистических признаков при обучении системы. Тем не менее, практические исследования показывают, что при отсутствии лингвистических признаков эффективный синтез интонации, особенно эмоциональной, возможен только при наличии огромного количества

интонационно разнообразного и представительного обучающего материала. Данная проблема затрагивает все языки, но становится еще более острой для тех из них, для которых доступно меньше качественных корпусов (в сравнении, например, с английским, на материале которого проводится большинство публикуемых исследований по синтезу).

По этой причине лингвистические признаки в системах синтеза все же не теряют своей актуальности и значимости: было показано, что, например, синтаксические признаки улучшают реализацию синтагматических границ (Guo et al., 2019), а использование эмбеддингов, учет пунктуации и частеречных тэгов повышает естественность интонационного контура (Tyagi et al., 2019). Учитывая данные результаты, можно предположить, что использование предсказанных интонационных моделей в разметке синтезируемого текста также внесло бы положительный вклад; это и обосновывает актуальность изучения методов предсказания интонационного оформления по тексту.

Практическая значимость работы заключается в том, что, во-первых, предлагаемый метод разработан на материале русского языка, для которого задача предсказания интонации исследована в значительно меньшей степени, чем, например, для того же английского; тем не менее, русский язык имеет свою специфику, в частности, используемые для него системы интонационного описания значительно отличаются от систем, обычно применяемых для английского. С этим связан второй аспект практической значимости работы: метод предсказывает интонационное оформление в терминах системы интонационного описания Н. Б. Вольской (Вольская и Скрелин, 2009), создание которой проводилось с учетом нужд и требований синтеза речи. Третий аспект практической значимости связан с тем, что метод использует нейросетевой подход, в последние годы позволивший значительно улучшить эффективность решения задач по обработке естественного языка.

Теоретической новизна состоит в изучении и сравнении эффективности предложенного метода на различном материале: на художественных и

новостных текстах, для дикторозависимого и дикторонезависимого материала, в условиях разнородности обучающей и тестовой выборок.

В качестве **материала** для исследования был использован речевой корпус CORPRES (COrpus of Russian Professionally REad Speech; Skrelin et. al, 2010), содержащий записи чтения, порожденные профессиональными дикторами. Аннотация корпуса включает просодический уровень, на котором отмечены различные интонационные явления: синтагматические границы, интонационные центры и модели и др. Наличие текстов, прочитанных всеми восемью дикторами корпуса, позволяет изучить пределы вариативности интонационного оформления; профессиональная подготовка дикторов обеспечивает уверенность в том, что предсказательная модель обучена на качественном материале, представляющем нормативную русскую речь.

Работа работа включает в себя введение, основную часть, разделенную на две главы, заключение, список литературы и приложение.

Первая глава содержит обзор литературы, на основе которой проводилась формализация признаков для предсказания интонационного оформления; описание используемых систем интонационного описания; обзор работ, посвященных предсказанию различных аспектов интонационного оформления.
Вторая глава содержит описание предлагаемого метода и его результатов: используемый для обучения материал, анализ вариативности предсказываемых интонационных явлений; классификационные признаки и архитектура модели; метрики; полученные результаты.

Теоретическая часть

1. Текстовые маркеры аспектов интонационного оформления

1.1. Функции интонации

Для того, чтобы понять, насколько теоретически возможно предсказание интонационного оформления высказывания на основе текстовых признаков, необходимо рассмотреть, какие именно функции выполняет интонация в речи и какими лингвистическими аспектами она обусловлена. Н. Д. Светозарова перечисляет следующие функции интонации (Светозарова, 1982):

- функция организации и членения речевого потока
- функция выражения степени связи между единицами членения
- функция оформления и противопоставления типов высказываний
- функция выражения отношений между элементами интонационных единиц
- функция выражения эмоциональных значений и оттенков

Последняя функция – эмоциональный аспект интонации – зачастую исключается из лингвистического анализа, поскольку не связана напрямую со смысловым содержанием сообщения, а обусловлена экстралингвистическими факторами, в частности, отношением говорящего к содержанию сообщения (Пронникова, 2014). Также предлагается называть данную функцию «экспрессивной», исключая из нее, таким образом, эмоциональный аспект (Кодзасов, Кривновна, 2001).

Остальные же функции имеют по большей части лингвистические основания: синтагматическое членение облегчает производство и восприятие речи с помощью членения текста на осмысленные отрезки (функция организации и членения речевого потока), что позволяет расставить акценты на наиболее важных частях высказывания; интонационное оформление синтагм

показывает, насколько сильно связаны между собой соседние синтагмы (функция выражения степени связи между единицами членения), каков коммуникативный тип каждой из них (функция оформления и противопоставления типов высказываний) и какова роль отдельных фонетических слов внутри синтагмы и их связь между собой (функция выражения отношений между элементами интонационных единиц). Таким образом, как синтагматическое членение, так и выбор интонационного оформления связаны с семантикой высказывания, поскольку способствуют выражению степени и характера связей между единицами текста на разных уровнях. Это позволяет предположить, что возможно с той или иной степенью точности предсказать допустимое интонационное оформление текста, исходя из его содержания, то есть, из его семантики, а, следовательно, и синтаксиса, являющегося отражением семантики. «Допустимое» интонационное оформление — это важная оговорка, поскольку высокая меж- и внутридикторская вариативность интонации заведомо делает нецелесообразной (в большинстве случаев) попытку предсказания единственно возможного интонационного оформления предложения.

Н. В. Черемисина-Ениколопова также отмечает, что синтагма является не только семантико-синтаксической единицей, но и «ритмо-интонационной, и стилистической» (Черемисина-Ениколопова, 1989), акцентируя внимание на том, что синтагматическое членение обусловлено не только семантически, но и ритмически, а стиль текста влияет на все аспекты интонационного оформления. Таким образом, данные параметры – ритмический и стилистический – также следует учитывать при предсказании интонационного оформления.

1.2. Синтагматическое членение и семантико-синтаксические факторы, обуславливающие его

Зависимость синтагматического членения от синтаксического членения является одним из наиболее активно обсуждаемых в научной литературе типов связи между интонацией и синтаксисом. Например, Э. Селкирк говорит о том, что границы синтаксических клауз в основном совпадают с границами

интонационных синтагм (intonational phrase; Selkirk, 2011). Теория М. Стидмана постулирует, что анализ синтагматического членения и синтаксической анализ предложений английского языка могут быть объединены в один процесс в системах автоматического понимания речи, а предписать высказыванию определенный тип интонационного акцента возможно, исходя из правил грамматики (Steedman, 1989), поскольку интонационная структура строго обусловлена фокусом и информационной структурой сообщения.

Многие лингвисты занимают менее радикальную позицию в вопросе соотношения синтаксической и интонационной структур. Например, С. Исихара спорит с Э. Селкирк, представляя случаи расхождений между синтаксическим и синтагматическим членением в японском языке и утверждая, что синтагматическое членение напрямую связано не с синтаксисом, а с прагматикой и дискурсивным членением (Ishihara, 2020).

Ряд принципов, контролирующих взаимодействие синтаксического и синтагматического членения, приводят М. Неспор и И. Фогель (Nespor, Vogel, 2008). Например, авторы говорят о том, что синтагматические границы в определенных синтаксических позициях более устойчивы и более частотны, чем другие. Например, наиболее устойчивыми являются границы между глагольными и именными группами (поскольку те являются рекурсивными синтаксическими узлами). Довольно вероятно появление границ между однородными членами предложения; с другой стороны, редки синтагматические границы между предикатом и его аргументами, а также внутри вводных конструкций, подтверждающих вопросов, обращений, восклицаний. Авторы, однако, отмечают, что ни одно из этих правил не является жестким предписанием, и может нарушаться при влиянии других факторов.

Тезис о большей «устойчивости» определенных синтаксических границ поддерживает, например, Н. А. Слюсарь, которая говорит о том, что с большей вероятностью в речи появляются те синтагматические границы, которые совпадают с наиболее «глубокими» синтаксическими границами, то есть, теми, что разделяют синтаксические составляющие, находящиеся на более глубоких

уровнях синтаксического дерева (Слюсарь, 2009). Автор объясняет варьирующуюся вероятность появления синтагматических границ в тех или иных местах синтаксической структуры тем, что последняя является более многоуровневой, чем интонационная, и «уплощается» в речи, теряя при этом наименее глубокие границы. Доля границ, существующих в синтаксической структуре предложения, но не имеющих интонационной манифестации, зависит от множества разнообразных факторов, например, от темпа речи.

Н. В. Черемисина-Ениколопова описывает отдельно как семантические, так и синтаксические факторы, обуславливающие синтагматическое членение (Черемисина-Ениколопова, 1989). К семантическим факторам автор относит следующие:

- цельность синтагмы с синтаксической точки зрения (т. е., синтагма должна объединять слова, принадлежащие к одной синтаксической группе);
- цельность синтагмы с семантической точки зрения (синтагма объединяет слова, связанные по смыслу, например, метафору и ключевые для ее понимания слова);
- антитеза (члены противопоставления зачастую разделяются синтагматической границей);
- степень знаменательности слов (знаменательные слова с большей вероятностью могут быть выделены в отдельную синтагму, в отличие от служебных или местоименных слов).

К синтаксическим факторам, обуславливающим синтагматическое членение, относятся такие факторы, как:

- сила синтаксической связи между словами: чем сильнее связь, тем меньше вероятность появления синтагматической границы между словами. К сильным синтаксическим связям относятся связи внутри предикативных, объектных, обстоятельственных, глагольных

атрибутивных синтаксических групп, именных атрибутивных групп с определениями и др. Данный фактор коррелирует с теми, что упоминались в работах Н. А. Слюсарь и М. Неспор и И. Фогель;

- порядок слов;
- наличие однородных членов;
- синтаксическая сочетаемость/несочетаемость связанных по смыслу слов;
- распространенность члена предложения. Например, распространенное подлежащее выделяется в отдельную синтагму в любых предложениях, а нераспространенное – в тех предложениях, где не распространено сказуемое. Другими примерами членов предложения, выделяемых в отдельную синтагму, могут послужить распространенное дополнение, нераспространенные препозитивные дополнение и обстоятельства, распространенное качественное обстоятельство, распространенное несогласованное определение, осложняющие структуры (вводные конструкции, обособленные обороты, в некоторых случаях – обращения) и др. При этом, данные правила являются, по мнению Н. В. Черемисиной-Ениколоповой, общеязыковыми, действующими для всех стилей речи, хотя и характеризующимися определенной вариативностью, которая зачастую приводит к отклонениям от описанных схем.

Исходя из сказанного в данном разделе, разумно сделать вывод, что синтагматическое членение высказывания обладает тесной связью с синтаксисом, которая, однако, не является абсолютной, и позволяет предсказывать расстановку синтагматических границ с ограниченной степенью точности.

1.3. Расположение интонационного центра синтагмы и его зависимость от синтаксических факторов

Определение позиции синтагматического ударения, или интонационного центра синтагмы, зачастую считается «автоматизированным» явлением

(Николаева, 1982): нормативным носителем синтагматического явления является последнее слово синтагмы. Известно, однако, также, что синтаксическая инверсия, будучи маркером «коммуникативной перестройки» (Николаева, 1982), приводит к изменениям в интонационной структуре, однако терминология, которую исследователи используют для того, чтобы описать происходящие изменения, разнится. Т. М. Николаева говорит о том, что при инверсии синтагматическое ударение, не значимое перцептивно для слушателя в силу того, что оно всегда находится в одной позиции, остается на последнем фонетическом слове предложения, однако член предложения, оказавшийся в инвертированной позиции, выделяется с помощью перцептивно значимого акцентного выделения.

И. И. Ковтунова же говорит не о двух различных по своей природе явлениях – центре синтагмы и дополнительном интонационном выделении, – а о двух видах фразового ударения: нейтральном (при отсутствии инверсии) и маркированном (при измененном порядке слов) (Ковтунова, 1976). Таким образом, согласно И. И. Ковтуновой, при инверсии происходит перенос ударения из одной позиции в другую, а не появление дополнительного акцента на инвертированном члене предложения. Автор также рассматривает правила немаркированного расположения ударения в различных синтаксических конструкциях (при анализе описанных правил можно увидеть, что ударение всегда падает на последнее знаменательное слово конструкции), например:

- в атрибутивных сочетаниях ударение падает на существительное;
- в глагольных сочетаниях – на зависимую от глагола словоформу;
- в словосочетаниях с качественными наречиями ударение падает на слово, к которому относится наречие;
- в составных сказуемых ударение падает не на вспомогательный глагол, а на другой член конструкции (инфинитив, именную часть, причастие, наречие).

Т. Е. Янко, хотя и не использует термин «синтагма», а связывает интонационное оформление предложения с его актуальным членением, также говорит о том, что можно формально вычислить носителя акцента как в теме, так и в реме, исходя из синтаксической структуры, фактора активации (известность/неизвестность референта для слушателя) и других признаков (Янко, 2008). К примеру, выбор носителя в реме определяется по следующим принципам:

- носителем акцента может быть только неактивированная, то есть, не известная из контекста синтаксическая группа;
- из оставшихся в результате применения первого правила синтаксических групп акцентоносителем становится та, что имеет наивысший приоритет в следующем ряду (синтаксические члены приведены в порядке возрастания приоритета):
  - Предикат → сирконстанты → актанты: A1, A2, A3, A4, A5, A6, …

В теме же акцента может не быть вовсе, а принцип активации не играет большой роли в выборе носителя, если он все-таки появляется.

Н. В. Черемисина-Ениколопова, называя расположение синтагматического ударения на последнем слове синтагмы законом русской речи, отмечает, что в экспрессивной, аффективной речи это правило нередко нарушается, место синтагматического ударения выбирается говорящим осмысленно, с опорой на контекст и смысл высказывания, а также закономерности расстановки ударений в разных типах синтаксических групп (Черемисина-Ениколопова, 1989). Синтагма с неконечным расположением синтагматического ударения обладает эмфатической динамической структурой, где возрастание силы ударений фонетических слов синтагмы происходит нелинейно; синтагма с конечным расположением центра обладает обычной динамической структурой, в которой сила ударений фонетических слов ступенчато возрастает по направлению к центру синтагмы.

Выделяются следующие ситуации, при которых часто наблюдается смещение синтагматических центров с последнего слова синтагмы:

- лексическая неполнозначность последнего слова синтагмы: при отсутствии контраста, незнаменательные слова редко становятся носителями синтагматического ударения независимо от их положения внутри синтагмы;
- инверсия;
- однородные члены в одном предложении или однотипные предложения. В этом случае ударение тяготеет к схожему компоненту предложений или однородным членам;
- наличие усилительных и ограничительных частиц – слово, к которому относятся такие частицы, притягивает синтагматическое ударение. Т. М. Николаева в данном случае также говорила о дополнительном акцентном выделении, а не переносе синтагматического ударения (Николаева, 1982);
- антитеза: помимо выделения членов противопоставления с помощью синтагматического членения, о чем говорилось выше, они также могут быть выделены с помощью переноса на них синтагматического ударения.

Автор отдельно отмечает тот факт, что имена существительные во всех стилях речи с большей вероятностью являются носителями акцента в то время как, например, прилагательное и глагол притягивают на себя интонационное выделение только при наличии определенных факторов – наличии противопоставления, ряда однородных членов или, в случае глагола, нераспространенного предложения с глаголом в конце.

Таким образом, можно предположить, что предсказание расположения интонационного центра синтагмы на основе синтаксических признаков возможно. В значительной степени это обусловлено тем, что при отсутствии инверсии или экспрессивной составляющей интонационный центр можно с высокой степенью надежности определить, опираясь лишь на расположение

синтагматических границ (которые, как было заключено в предыдущем разделе, можно предсказать из синтаксического членения) и на информацию о знаменательности слов синтагмы: как правило, носителем синтагматического ударения оказывается последнее знаменательное слово синтагмы. Инверсия, которая приводит к смещению интонационного центра (или появлению дополнительного интонационного выделения), также может быть определена с помощью автоматического синтаксического анализа.

1.4. Факторы, влияющие на выбор интонационного оформления синтагмы

Факторы, обуславливающие интонационное оформление синтагмы, изучалась такими отечественными лингвистами, как, например, О. Ф. Кривновой (Кривнова и др., 2016), Н. Д. Светозаровой (Светозарова и Штерн, 1989), Н. Б. Вольской (Вольская и Скрелин, 2009). Однако, в данном разделе будут, в первую очередь, рассмотрены признаки, формулируемые Т. Е. Янко (Янко, 2008), поскольку они подробно характеризуют именно влияние синтаксиса, лексики и прагматики на аспекты интонационного оформления.

Как отмечает Т. Е. Янко, особенность интонации, в отличие от остальных уровней языка, заключается в том, что ее план содержания крайне абстрактен – значения, передаваемые интонацией, размыты и трудно поддаются формулировке, как за счет своей непредметности, так и за счет широко распространенной омонимичности. Тем не менее, начиная анализ интонационных стратегий русской речи, Т. Е. Янко выделяет основные типы значений, выражаемых интонационными средствами:

- иллокутивные – значения темы и ремы; среди них выделяются следующие подтипы:
    - системные, задающие иллокутивную грамматику языка, такие, например, как интонация вопроса или повествования. Могут быть дополнены модифицирующими значениями, не меняющими их

основного значения. Автор считает системные значения универсальными для большинства языков. К ним относятся:

- собственно иллокутивные (рема);
- несобственно иллокутивные (тема);

◦ уникальные (словарные), различающиеся между языками и являющиеся «монолитными структурами», поскольку они не подвергаются модификациям и зачастую даже не имеют конкретного интонационного центра. В русском языке к таким относятся, например, интонация безрезультатной деятельности или зова невидимого адресата;

- модифицирующие значения, накладываемые на первичные иллокутивные значения:
  ◦ эмфаза;
  ◦ контраст;
    - собственно контраст (в данном случае множество противопоставляемых объектов насчитывает более двух альтернатив);
    - верификация, частный случай контраста (выбор производится из двух альтернатив: да/нет);
- собственно дискурсивные, отвечающие за связь между предложениями в тексте. К таким, например, относится интонация незавершенности;
- вторичные иллокуции, передающие речевые акты мольбы и угрозы и выражающие состояние говорящего. Они манифестируются такими

интонационными средствами, как тембр, усиленная интенсивность, четкость произнесения, а также не принадлежат одной синтагме, выходя даже за рамки предложения и действуя иногда на протяжении целого текста.

Т. Е. Янко также отмечает, что значения темы, ремы, вопроса (иллокутивные значения), незавершенности (дискурсивные значения) маркируются различными типами интонационных акцентов, в то время как модифицирующие значения меняют лишь интенсивность интонационного акцента, а эмфаза приводит к искривлению основного тона (то есть, данные значения обусловливают вариативность среди акцентов одного и того же типа). Автор также приводит примеры связи конкретных значений и интонационных контуров (типы акцентов обозначаются в соответствии с интонационной системой русского языка, разработанной Е. А. Брызгуновой (Брызгунова, 1963)):

- ИК-2 и ИК-3 являются показателями темы, при этом ИК-3 является немаркированным, нейтральным вариантом, а ИК-2 служит для дополнительного интонационного выделения;
- ИК-3 и ИК-6 являются взаимозаменяемыми маркерами незавершенности только для темы;
- в контрастной теме после интонационного центра обычно наблюдается нисходящее движение ОТ;
- эмфаза выделяется с помощью тонального перепада: к изначальному движению основного тона присоединяется движение, направленное в противоположную сторону (к восходящему – нисходящее, и наоборот).

Н. В. Черемисина-Ениколопова рассматривает десять типовых мелодических контуров, соответствующих разным коммуникативным значениям (Черемисина-Ениколопова, 1989). К ним относятся:

- констатирующая, передающая значение завершенности, характеризуемая нисходящим движением основного тона и встречающаяся в синтагмах почти всех синтаксических типов, причем наиболее часто – в случае финальной синтагмы предложения;
- восклицательная, также характеризуемая нисходящим движением, но с более высокого уровня, а также большей интенсивностью и длительностью. Восклицательные мелодемы по звучанию схожи с констатирующими мелодемами с дополнительной эмфазой на каком-либо из членов синтагмы;
- вопросительная, с повышением основного тона на вопросительном слове или, при его отсутствии, слове, содержащем основной смысл вопроса;
- начинательная, с ровным тоном в среднем регистре в основной части синтагмы и восходящим движением в центре; характерна для начальных синтагм предложения вне зависимости от их состава;
- перечислительная, реализуемая на цепочке однородных компонентов, выделенных в отдельные синтагмы. Акустическое оформление может быть разным: как восходящим или нисходящим, так и восходяще-нисходящим;
- противительная, имеющая восходящую «вогнутую мелодическую кривую», завершающуюся нисходящим движением. Реализуется при наличии противопоставления;
- пояснительная, встречающаяся при выделении в отдельную синтагму члена предложения, отделенного на тексте двоеточием. Имеет нисходящее в среднем или низком регистре движение основного тона, ускоренный темп;
- императивная, свойственная обращению и предложениям-распоряжениям. Выражается падением тона с высокого уровня. Эмфатический вариант, реализуемый на двусловном обращении,

характеризуется переносом синтагматического ударения на первое слово синтагмы;

- вводная, реализуемая на вводных конструкциях и других отрезках речи, не имеющей больной важности; характеризуется слабым, монотонным понижением тона, сниженной громкостью и ускоренным темпом;
- уточняющая, типичная для обособленных членов предложения. Обычно имеет восходящий интонационный контур.

В речи возникают варианты перечисленных мелодем, как обусловленные позицией синтагм в тексте, так и сочетанием мелодем соседствующих синтагм. На стандартные контуры могут накладываться эмоционально-экспрессивные модификации. Эмоциональный пик текста, приводящий к появлению более ярких и интенсивных с точки зрения разных аспектов вариантов мелодем, зачастую притягивается к тропам и лексическим единицам определенных типов, например:

- сравнениям;
- эмоционально-стилистически окрашенным словам;
- словам, обозначающим эмоции;
- эмоционально повторяемым словам;
- неологизмам;
- словам, употребленным в переносном или гиперболическом значении.

Опираясь на наблюдения Т. Е. Янко и Н. В. Черемисиной-Ениколоповой, можно формализовать некоторые факторы, обуславливающие выбор интонационного оформления синтагмы, чтобы в дальнейшем использовать их в качестве классификационных признаков.

2. Система интонационного описания

Данная работа фокусируется на предсказании интонационного оформления высказывания в терминах системы интонационных моделей, разработанной Н. Б. Вольской (Вольская и Скрелин, 2009). Выбор обоснован тем, что данная система была создана для применения в области синтеза речи по тексту. В отличие от классификации мелодем Н. В. Черемисиной-Ениколоповой, приведенной в разделе 1.4, в системе интонационного описания Н. Б. Вольской типы интонационных моделей выделяются не только на основе значений, передаваемых ими, но и на основе их акустических различий. Таким образом, интонационные единицы, используемые для передачи сходных значений, но имеющие систематически различающиеся планы выражения, выделяются в данной системе в две различные интонационные модели. Основой системы Н. Б. Вольской служит система интонационного описания русского языка Е. А. Брызгуновой (Брызгунова, 1963), однако система Н. Б. Вольской расширяет инвентарь интонационных моделей за счет добавления тех, что не были рассмотрены в системе Е. А. Брызгуновой, а также за счет разделения на несколько моделей интонационных единиц, имеющих одинаковый план выражения, но разные значения.

Как и в системе Е. А. Брызгуновой, каждая синтагма, согласно системе описания Н. Б. Вольской, делится на три части: центр, совпадающий с ударным слогом слова, на которое падает синтагматическое ударение, и выделенный наиболее значимым движением основного тона; предцентр, предшествующий центру, и постцентр, который может отсутствовать в том случае, если ударный слог слова с синтагматическим ударением является последним в синтагме. В таблице 1 представлены интонационные модели, выделяемые в системе Н. Б. Вольской, их значения и соответствия в системе Е. А. Брызгуновой.

Таблица 1. Система интонационных моделей Н. Б. Вольской (Вольская и Скрелин, 2009; Volskaya & Kachkovskaia, 2016).

Номер модели по системе Н. Б. Вольской	Номер модели по системе Е. А. Брызгуновой	Коммуникативный тип/значение	Примечания
01	1	завершенность	полная завершенность (напр., в конце абзаца)
01a	1	завершенность	наиболее частотный вид завершенности, как правило, на конце предложений
01b	1	завершенность	неполная завершенность, напр., внутри предложения
02	2	выделенность	Стандартная реализация ИК-2 (система Е. А. Брызгуновой)
02b	2	выделенность	сложный восходяще-нисходящее-восходяще-нисходящий контур
02c	2	выделенность	падение с очень высокого уровня в низкий регистр

03	2	вопрос	интонационный центр на вопросительном слове
03a	2	вопрос	интонационный центр не на вопросительном слове
04	2	восклицание	
04a	2	обращение	
04b	2	просьба	
05	5	восклицание	с дополнительным интонационным выделением внутри синтагмы
06	1/6	восклицание	низкий ровный тон
06a	6	восклицание	высокий ровный тон
06b	6	вопрос	вопрос-уточнение
06c	6	вопрос	вопрос с оттенком недоумения
07	3	вопрос	интонационный центр на последнем слове

07a	3	вопрос	интонационный центр не на последнем слове
07b	3	вопрос	мелодический пик сдвинут на постцентр
08	4	вопрос	
08a	4	вопрос	экспрессивная окраска
09	--	комментарий	нисходящий тон
09a	1	комментарий	ровный тон
09b	3/4/6	комментарий	восходящий тон
10	1	незавершенность	
11	3	незавершенность	
11a	3	незавершенность	мелодический пик сдвинут на постцентр
11b	3	незавершенность	с дополнительным интонационным выделением внутри синтагмы
12	6	незавершенность	
12a	--	незавершенность	ровный тон в среднем регистре
13	4	незавершенность	

3. Обзор работ по предсказанию интонационного оформления текста

Предсказание интонации по информации, содержащейся в тексте, является одним из ключевых моментов синтеза речи по тексту. Корректное предсказание синтагматического членения и качественное выполнение интонационного оформления высказывания является залогом естественно звучащей, легко воспринимаемой синтезированной речи, верно передающей смысловое содержание сообщения. Данная задача, однако, нетривиальна вследствие того, что все аспекты интонации характеризуются чрезвычайно высокой вариативностью, зависящей как от лингвистических параметров текста (например, его жанра), так и от особенностей конкретного диктора, его стиля речи и нюансов его личной интерпретации текста. При моделировании интонации в системе синтеза речи необходимо учитывать наличие данной вариативности: с одной стороны, необходимо разнообразить интонационное оформление синтезируемой речи для того, чтобы приблизить ее по натуральности к человеческой; с другой стороны, важно осознавать пределы, в которых можно варьировать интонационное оформление без искажения содержания. Как утверждает П. Тэйлор в своей книге, посвященной синтезу речи по тексту (Taylor, 2009), для предсказания особенно сложен эмоциональный компонент интонации, поскольку он, как правило, никак не закодирован в самом тексте, и добавляется уже непосредственно самим диктором при прочтении. Повторить это невозможно без понимания содержания текста — операции, попытки осуществления которой в современных системах синтеза речи не производятся. Тем не менее, это не мешает синтезу речи широко применяться в тех областях, где моделирование эмоционального компонента интонации не востребовано: например, в автоинформационных системах, call-центрах, в системах звукового оповещения, для озвучивания текстов для слепых и слабовидящих людей и так далее (Рыбин, 2014).

3.1. Постановка задачи при предсказании интонационного оформления

В связи с постоянным развитием систем синтеза, задача по созданию и совершенствованию алгоритмов предсказания интонационного оформления текста является предметом тщательного изучения с 90-х годов прошлого века (см., например, (Hirschberg, 1991; Ostendorf et al., 1993)) и не теряет своей актуальности по сей день (например, (Sloan et al., 2019)). Как правило, работы в данной области фокусируются на предсказании положения границ интонационных единиц (prosodic breaks) в тексте и/или положения тональных акцентов (pitch accents) и их классификации; небольшое число исследований нацелено на предсказание эмфатически выделенных слов (Brenier et al., 2005; Nakajima et al., 2014) или классификации пограничных тонов (Rendel et al., 2016). Наибольший успех был достигнут в задаче по предсказанию положения границ интонационных единиц, где F1-мера в большинстве случаев выше 80% (Mishra et al., 2015; Sloan et al., 2019; Tepperman & Nava, 2011), а для некоторых языков превышает 90% (Liu et al., 2018; Menshikova & Kocharov, 2019); тем не менее, стоит отметить, что результаты чрезвычайно сильно зависят от конкретного корпуса, используемого для обучения предсказательной модели, и типа речи, представленного в нем (чтение или спонтанная речь), а также от того, границы каких именно интонационных единиц подлежат предсказанию: в исследовании, проведенном на материале французского языка, было показано, что, в то время, как предсказание границ более крупных единиц (major phrase breaks) находилось на уровне 95% F1-меры, точность предсказания границ между более мелкими интонационными единицами (intermediate phrase breaks) падала практически вдвое, до 55% F1-меры (Obin & Lanchantin, 2015). Задача по предсказанию положения и типов тональных акцентов представляется более сложной, поскольку характеризуется еще большей, чем расстановка синтагматических и фразовых границ, вариативностью, за счет менее тесной связи с такими поверхностными, легко извлекаемыми признаками, как пунктуация, и за счет отсутствия физиологических ограничений: поскольку

внутри синтагмы невозможны дыхательные паузы (Светозарова, 1982), длина синтагм в некоторой мере ограничена необходимостью совершать вдох. Расстановка тональных акцентов же не ограничена явлениями подобного рода, универсальными для всех людей, и сильнее варьируется от языка к языку. Тем не менее, в предсказании положения тональных акцентов для английского языка на данный момент стабильно преодолевается планка в 80% (например, 84% F1-меры в (Sloan et al., 2019)). Что касается предсказания эмфатически выделенных слов, то работы, представленные для данной задачи, немногочисленны и сильно разнятся между собой как по постановке эксперимента (наличие/отсутствия предположения о том, что в каждом предложении есть эмфатически выделенное слово), так и по полученным результатам: от 31% F1-меры в (Brenier et al., 2005) и 44% в (Hovy et al., 2013) до 67% в (Nakajima et al., 2014).

3.2. Связь между предсказанием интонационных границ и тональных акцентов

Зачастую в исследованиях, рассматривающих одновременно методы предсказания синтагматических границ и тональных акцентов, две эти задачи оказываются не связаны между собой: модели для каждой из них обучаются отдельно друг от друга, и результаты, полученные в ходе предсказания синтагматических границ, никак не используются при предсказании/классификации тональных акцентов (например, подобная постановка эксперимента наблюдается в (Obin & Lanchantin, 2015; Sloan et al., 2019; Chen et al., 2004; Tepperman & Nava, 2011; Rendel et al., 2016)), несмотря на то, что, представляется, что эти аспекты интонационного оформления взаимосвязаны и заметно влияют друг на друга. Вероятно, подобная тенденция обусловлена тем, что в зарубежных интонационных исследованиях широко распространена система ToBI (Silverman et al., 1992), в терминах которой осуществляется просодическая разметка наиболее известных и репрезентативных речевых корпусов, используемых для обучения значительной части предсказательных моделей (такими корпусами для английского языка

являются, например, корпус радио-новостей, созданный в Бостонском университете (Ostendorf et al., 1995), или игровой корпус, созданный в Колумбийском университете (Gravano & Hirschberg, 2011)). В рамках системы ToBI интонационное оформление высказывания интерпретируется как цепочка интонационных событий — тональных акцентов, пограничных тонов или границ интонационных единиц различных уровней, при этом количество тональных акцентов, которые могут находиться внутри одного высказывания или интонационной группы, не ограничено, а их положение, соответственно, не считается тесно связанным с интонационными границами. В противоположность подобному взгляду на интонацию, в отечественной традиции, например, в интонационной системе Е. А. Брызгуновой (Брызгунова, 1963) и в системе Н. Б. Вольской (Вольская и Скрелин, 2009), каждая минимальная интонационная единица — синтагма — рассматривается как носитель единого мелодического контура, имеющего конфигурацию одного из заранее определенных типов, а также четкую структуру, состоящую из центра, предцентра и постцентра, при этом центр, как правило, привязан к ударному слогу последнего фонетического слова синтагмы. Таким образом, в системах описания интонации русского языка связь между мелодическим контуром и синтагмой (соответственно, ее границами) прослеживается более отчетливо, чем связь между цепочкой тональных акцентов и границами интонационных единиц в ToBI, и потому можно предположить, что для предсказания интонационного оформления русского текста разумно было бы организовать эксперимент как минимум в два последовательных этапа: предсказание синтагматических границ и, затем, предсказание интонационных центров найденных синтагм. С другой стороны, важно отметить, что верхний порог возможной точности предсказания становится тем ниже, чем более емкая и сложная по структуре единица рассматривается, и потому стоит ожидать, что при прочих равных условиях результаты по предсказанию типа интонационной конструкции будут хуже, чем результаты по предсказанию типов тональных акцентов (то есть, движений основного тона, привязанных к одному слову, а не

к целой синтагме). Также потребуется разработка методов анализа признаков, характеризующих всю синтагму в целом: в исследованиях, существующих на данный момент, самой крупной единицей, для которой извлекаются признаки, применяемые для предсказания интонационного оформления, является орфографическое слово (однако, зачастую к этим признакам добавляются контекстные признаки — признаки соседних слов, что увеличивает точность предсказания расположения тональных акцентов (Sloan et al., 2019)).

3.3. Зависимость точности предсказания от материала, используемого для обучения

Особое значение в контексте предсказания интонационного оформления по тексту имеют исследования, рассматривающие применимость моделей, обученных на одном корпусе, для использования на материале другого. Работа Р. Слоан и ее коллег (Sloan et al., 2019) показала, что модели для предсказания синтагматических границ лишь незначительно теряют в точности, если их обучать и применять на разных корпусах, в которых представлена речь одного типа: при тестировании модели, обученной на корпусе Бостонского университа, на другом корпусе, содержащем записи чтения новостных текстов, F1-мера упала на 1,1% для предсказания положения тональных акцентов и на 5,1% для предсказания синтагматических границ по сравнению с результатами, полученными при тестировании на материалах Бостонского же корпуса. В случае значительного изменения стиля речи в тестовом материале по сравнению с обучающим, разница в точности, как и следует ожидать, становится слишком большой, чтобы ей можно было пренебречь: 14,6% ухудшения в точности для предсказания тональных акцентов и 34,1% — для синтагматических границ в том случае, если модель была обучена на материале чтения и протестирована на спонтанной речи; 0,04% и 42,8% для модели, обученной на спонтанной речи и протестированной на чтении (малое ухудшение точности в предсказании тональных акцентов связано с тем, что модель, обученная на спонтанной речи, демонстрировала крайне низкую точность, даже когда применялась на материале того же типа).

В ходе исследования Н. Обена и П. Ланшантена (Obin & Lachantin, 2015) было сделан вывод о слабой применимости моделей, обученных на материале спонтанной речи, к материалу чтения, особенно в том случае, если модель использует синтаксические классификационные признаки: точность синтаксического анализа для спонтанной речи значительно ниже, от чего страдает и качество предсказательной модели.

Однако, А. Розенберг и др. (Rosenberg et al., 2012) утверждают, что при тестировании моделей на материале различных типов, выяснилось, что предсказание синтагматических границ в новостных и монологических текстах производится моделью, обученной на спонтанных диалогах, с более высокой точностью, чем наоборот.

Тем не менее, можно утверждать, что с высокой степенью надежности можно применять предсказательные модели для синтеза речи лишь того типа, что был представлен в обучающем материале данной модели; смешение спонтанной и подготовленной речи всегда приводит к значительному ухудшению результатов.

3.4. Классификационные признаки

Как правило, в исследованиях, рассматривающих одновременно и задачу по предсказанию синтагматических границ, и задачу по предсказанию тональных акцентов, все извлекаемые из текста классификационные признаки используются как для обучения первой модели, так и для обучения второй (ссылки на подобные исследования даны выше); то есть, не проводится граница между признаками, которые могут быть полезны исключительно для определения интонационных границ или для классификации движений основного тона. Схожим образом, в данном разделе не будут противопоставляться признаки, которые применяются для той или иной задачи по предсказанию интонационного оформления; все они будут приведены единым списком.

Предполагается, что выбор интонационного оформления высказывания в значительной степени обусловлен семантикой высказывания; однако доступные

на данный момент инструменты для автоматического семантического анализа не характеризуются большим разнообразием и высоким уровнем точности, особенно для языков, отличных от английского. Более того, формализация семантики высказывания таким образом, который позволил бы использовать ее в качестве классификационных признаков для статистической модели представляет собой отдельную нетривиальную задачу. Поэтому в современных методах предсказания интонационного оформления по тексту основную роль играют грамматические, лексические и синтаксические признаки.

Наиболее распространенным признаком являются частеречные тэги (Mishra et al., 2015; Kocharov et al., 2019; Read & Cox, 2007; Schmid & Atterer, 2004; Louw & Moodley, 2016; Chen et al., 2004; Obin & Lachantin, 2015; Rendel et al., 2016; Klimkov et al., 2017; Hirschberg et al., 1996; Ingulfsen, 2004): во-первых, частеречная разметка в некоторой степени косвенно отражает синтаксическую структуру высказывания, но при этом отличается большей надежностью, чем автоматический синтаксический разбор, особенно при анализе орфографических расшифровок спонтанной речи, где точность синтаксических парсеров падает; во-вторых, частеречные тэги сужают круг слов, которые могут оказаться носителями тонального акцента. В моделях, не учитывающих контекст классифицируемого объекта, используются не отдельные тэги, а частеречные N-граммы (Sun & Applebaum, 2001). В работе В. Климкова и его коллег (Klimkov et al., 2017) исследовалась возможность замены частеречных тэгов признаком, указывающим на принадлежность лексической единицы к служебным либо знаменательным словам: как было показано, это имеет смысл (как минимум на материале английского языка), поскольку качество системы не становилось хуже в сравнении с системами, использующими полный набор частеречных тэгов.

Совместно с частеречной разметкой зачастую применяются разнообразные синтаксические признаки. Их роль считается особенно важной при предсказании синтагматических границ: так, (Koehn et al., 2000) и (Obin & Lachantin, 2015) сообщают, что синтаксическая информация позволяет добиться

значительного повышения точности предсказания, а в (Mishra et al., 2015) комбинация частеречных тэгов и синтаксических признаков позволила добиться того же уровня точности, который дали лексические признаки (при этом, как представляется, модель, натренированная на лексических признаках, в отличие от грамматических и синтаксических, будет менее надежна при применении на любом другом корпусе). С другой стороны, в эксперименте, проведенном на материале русского языка, добавление синтаксических признаков к частеречным тэгом позволило получить прирост F1-меры лишь на 1% (Menshikova & Kocharov, 2019); а в работе, рассматривающей вклад синтаксических признаков, составленных на основе различных грамматик, в предсказание синтагматических границ, указывается, что признаки, полученные в результате самого простого и поверхностного синтаксического анализа, оказались примерно настолько же результативны, как и более емкие и глубинные синтаксические представления, на извлечение которых необходимо затратить больше времени (Ingulfsen, 2004). Что же касается предсказания тональных акцентов, то (Obin & Lachantin, 2015) отмечают, что глубинная синтаксическая информация не приносит существенной пользы при решении данной задачи (авторы предполагают, что это происходит потому, что тональные акценты расставляются в соответствии с семантическим фокусом, который не отражается в содержании синтаксических признаков).

Представление синтаксических признаков во многом зависит от того, какая грамматика используется при синтаксической разметке материала — грамматика зависимостей или грамматика составляющих. В первом случае в качестве признаков может использоваться непосредственно тэг, обозначающий синтаксические отношения, существующие между текущим словом и тем, от которого оно зависит (Kocharov et al., 2019; Klimkov et al., 2017); существование зависимости между текущим словом и соседними (Kocharov et al., 2019); бинарный признак, указывающий на то, имеются ли у данного слова зависимые; количество зависимых слов; максимальное расстояние от текущего слова до зависимого от него (Mishra et al., 2015); синтаксический тэг слова, главного к

текущему (Klimkov et al., 2017); длина кратчайшего пути от узла текущего слова до корня предложения. В случае грамматики составляющих чаще используются признаки, извлеченные из более «глубоких» уровней синтаксического дерева: типы группы, в которую входит текущее слово (Obin & Lachantin, 2015); глубина дерева для текущего слова (Klimkov et al., 2017); длина самой крупной группы, которая оканчивается данным словом; длина следующей группы, находящейся на том же уровне синтаксического дерева; количество дочерних узлов у самой крупной группы, которая оканчивается данным словом; ее тип; длина пути по синтаксическому дереву, позволяющая достигнуть следующее слово в предложении (Read & Cox, 2007); ширина и глубина самого маленького дерева, содержащего текущее слово, ширина и глубина корневого узла синтаксического дерева, глубина наименьшего дерева, содержащего текушее слово и следующее за ним (Sloan et al., 2019); количество синтаксических групп, начинающихся с текущего слова, и количество групп, заканчивающихся им (Chen et al., 2004); бинарный признак, показывающий, находится ли окно из N-го количества слов, в которое входит текущее слово, внутри именной группы или рядом с ней, и если находится, то какова длина данной именной группы (Hirschberg et al., 1996); длина следующей группы, находящейся на том же уровне синтаксического дерева, как и самая крупная группа, содержащая данное слово (Ingulfsen, 2004).

Распространены и различные виды лексических признаков, которые помогают, в частности, предсказывать лексические единицы, с наибольшей вероятностью притягивающие к себе тональный акцент. Например, могут применяться леммы слов (Mishra et al., 2015), однако отмечается, что данный признак ненадежен, поскольку частотность лемм сильно изменяется в зависимости от жанра и тематики текста. В работе Рэндела и др. (Rendel et al., 2016) для предсказания интонационного оформления использовались шесть бинарных признаков, указывающих на то, является ли текущее слово: а) написанным с заглавной буквы; б) предлогом или послелогом; в) союзом; г) вспомогательным глаголом; д) вопросительным словом; е) служебным словом.

В другом исследовании, где применялись те же классы, за исключением первого (Rosenberg et al., 2012), был сделан вывод, что данные признаки, как и некоторые синтаксические, обладают заметной чувствительностью к материалу, и при смене его типа становятся малоинформативны (вследствие того, что синтаксическая структура и лексическое наполнение разных типов речи сильно отличаются друг от друга). В этой же работе применялся т.н. коэффициент акцентуации (‘accent ratio’): статистически-лексический признак, показывающий, как часто данная лемма оказывалась носителем тонального акцента в обучающем материале.

Последнее время пользуется популярностью такой признак, как эмбеддинг (векторное представление) слова: в исследовании (Klimkov et al., 2017), посвященном предсказанию синтагматических границ, система с единственным классификационным признаком — эмбеддингом слова — продемонстрировала результаты, близкие к результатам системы, использующей богатый набор синтаксических признаков. В то же время, по итогам работы (Rendel et al., 2016), где были протестированы 3 различных метода векторного представления слов, был сделан вывод, что для предсказания синтагматического членения эмбеддинги представляют малую ценность (F1-мера повысилась на 1,9% после применения эмбеддингов); также авторы заключили, что выбор размерности эмбеддингов играет большую роль, чем выбор метода обучения векторной модели. Для монгольского языка с помощью эмбеддингов морфем, то есть, более мелких, чем слово, единиц, был продемонстрирован вариант решения проблемы, связанной со словами, отсутствующими в обучающей выборке (Liu et al., 2018). В статье Слоан и др. (Sloan et al., 2019), помимо кластеризованных эмбеддингов отдельных слов, в качестве признаков также были использованы векторные представления целых предложений, которые оказались более надежды и эффективны за счет богатой контекстуальной информации, заключенной в них.

Та же работа Слоан и др. рассматривает множество признаков, ранее не использовавшихся для предсказания интонационного оформления:

- тэги именованных сущностей текущего слова и следующего за ним;
- признаки, извлекаемые с помощью инструмента LIWC (Linguistic Inquiry and Word Count; (Pennebaker et al., 2015)) и показывающие, принадлежит ли лемма к каким-либо из 73 заранее выделенных абстрактных категорий слов;
- кореференциальные признаки: число упоминаний слова ранее в текущем тексте, расстояние от текущего слова до его последнего употребления, синтаксическая функция его последнего употребления, частеречный тэг и синтаксическая функция последнего имплицитного упоминания текущего слова;
- признак, извлекаемый с помощью инструмента Speciteller (Li & Nenkova, 2015): коэффициент конкретности текущего предложения, значение которого варьируется в пределах от 0 (предложение, содержащее наиболее общую информацию) до 1 (предложение, содержащее максимально конкретную информацию).

Из этих признаков наиболее информативными оказались признаки LIWC, особенно для служебных слов, и коэффициент Speciteller.

Широко распространено использование дистанционных признаков, характеризующих длину различных частей предложения или текста: длина предложения, позиция слова в текущем предложении (Sloan et al., 2019), расстояние до предыдущего знака пунктуации (Mishra et al., 2015), количество слов и слогов в предложении, расстояние до конца предложения (Hirschberg et al., 1996). Так как синтагматическое членение в значительной степени обусловлено расстановкой знаков препинания, последние также могут использоваться в качестве отдельных признаков.

3.5. Классификационные методы

Методы, использующие вышеприведенные признаки для предсказания интонационного оформления текста, также отличаются разнообразием. Это могут быть системы, построенные на правилах (Лобанов, 2008), однако

большей популярностью отличаются методы машинного обучения, среди которых условные случайные поля (Conditional Random Fields; Kocharov et al., 2019; Louw & Moodley, 2016; Rosenberg et al., 2012), деревья классификации и регрессии (Classification and Regression Trees) (Helander & Nurminen, 2007; Hirschberg et al., 1996; Hirschberg, 1993), скрытые марковские модели (Hidden Markov Model; Obin et al., 2015), случайные леса (Random Forests; Sloan et al., 2019), деревья решений (Bulyko & Ostendorf, 2001). В последнее время набирают популярность методы глубинного обучения: например, рекуррентные нейронные сети (Pascual & Bonafonte, 2016), двунаправленные рекуррентные нейронные сети (Rendel et al., 2016; Rendel et al., 2017), двунаправленные LSTM (Liu et al., 2018), эффективность которых по сравнению с другими методами машинного обучения была доказана на материале английского (Klimkov et al., 2017) и русского языков (Menshikova & Kocharov, 2019). Выбор метода классификации зависит от условий конкретного эксперимента: например, более трудоемкие и ресурсозатратные методы, такие, как нейросети с долгой краткосрочной памятью (LSTM), способны лучше простых статистических методов смоделировать сложные зависимости признаков, но требуют большого объема обучающего материала; текущих же объемов речевых корпусов, особенно, просодически размеченных, не всегда оказывается достаточно. Для того, чтобы уменьшить объем необходимого обучающего материала, можно уменьшить количество используемых признаков (например, снижать размерность векторных представлений или снижать число контекстных признаков — как показало исследование (Mishra et al., 2015), модель, учитывающая только левый контекст слова, почти так же точна, как модель, использующая и правый, и левый). Методы, основанные на частичном (в отличие от полного) обучении модели с учителем (weakly/fully supervised models), также помогают бороться с нехваткой обучающего материала (Rendel et al., 2017).

Практическая часть

1. Материал

1.1. Описание корпуса

В работе использовались материалы корпуса CORPRES (COrpus of Russian Professionally REad Speech; Skrelin et. al, 2010), созданного на кафедре фонетики и методики преподавания иностранных языков Санкт-Петербургского государственного университета. Корпус состоит из записей чтения восемью профессиональными дикторами, суммарная продолжительность которых составляет около 30 часов. Материал для чтения содержал художественные повести, пьесу, новостные тексты на политические и экономические темы, а также нейтральные тексты про IT сферу. Аннотация корпуса включает несколько уровней, в частности, орфографическую расшифровку и просодический уровень, на котором указаны границы синтагм, интонационные модели, интонационные центры, типы пауз и эмфатические выделения.

Поскольку интонация при чтении текстов разных жанров значительно варьируется (Бондарко и др., 1988), было решено использовать для обучения системы тексты одного жанра. При этом, большая часть экспериментов проводилась на материале художественных повестей (А. Г. Алексин, «Поздний ребенок», и Ю. В. Трифонов, «Обмен»), поскольку они представлены в корпусе наиболее полным образом (каждую из них прочитали все восемь дикторов), а также поскольку они обладают большим разнообразием коммуникативных типов предложений и, как следствие, большим интонационным разнообразием (по сравнению с новостными и информационными текстами), в частности, за счет наличия в них эмфатических интонационных моделей. Однако был также проведен ряд дополнительных экспериментов для новостных текстов, прочитанных двумя дикторами.

Художественные повести содержат 3874 предложения и более чем 35000 слов. Около 80000 синтагм было реализовано всеми восемью дикторами корпуса при чтении данных текстов, в среднем — около 10000 синтагм на диктора. Новостные тексты включали 1849 предложений, насчитывающих более чем 27000 слов и около 15000 синтагм, порожденных двумя дикторами, в среднем — около 7500 синтагм на диктора.

1.2. Описания обучающих и тестовых выборок

Выборки для обучения и тестирования модели состояли из отдельных предложений. Предложения были перемешаны случайным образом (таким образом, структура текста не была сохранена, хотя она частично учитывалась в признаках). Предложения были токенизированы с помощью модуля DeepPavlov (Burtsev et al., 2018); перед подачей в нейросеть из предложений были удалены знаки препинания. Разбиение на обучающую и тестовую выборку проводилось в соотношении 75:25.

Для обучения и тестирования моделей по разным параметрам было отобрано несколько выборок:

1. Выборка для тестирования параметров нейросетевой модели и выбора их наилучших значений. Обучающая выборка являлась частью подкорпуса художественных текстов (были включены обе повести и прочтения всех восьми дикторов) и составляла 75% от него. Предложения, входящие в обучающую выборку, были выбраны случайным образом без учета текста, в который они входили, диктора, породившего их, и того, входила ли другая реализация предложения в тестовую выборку. Тестирование моделей проводилось на оставшихся 25% подкорпуса (тестирование проводилось единожды, без кросс-валидации).
2. Интонационно-сбалансированная выборка. Был проведен эксперимент, направленный на изучение того, может ли учет интонационных моделей, употребленных диктором в предложении, при составлении обучающей выборки улучшить результаты. Выборка составлялась схожим образом с

выборкой из п. 1 (без кросс-валидации, учета дикторов и предложений), однако контролировалось количество употреблений каждой интонационной модели в обучающей выборке: их должно было быть 75% от общего числа. Оставшиеся 25% входили в тестовую выборку.

3. Общая выборка для дикторонезависимой модели (художественные тексты). Выборка включала обе повести и всех дикторов, однако в данном случае проводилась кросс-валидация по диктору: было составлено 8 пар обучающих и тестовых выборок; в тестовую выборку было включено 25% предложений одного из дикторов корпуса, а обучающая выборка состояла из предложений, порожденных семью остальными дикторами корпуса. При этом, в обучающую выборку было включено 75% не задействованных в тестовой выборке предложений, чтобы исключить пересечение между предложениями обучающей и тестовой выборок. Для каждой пары выборок была обучена модель, а тестовые показатели всех полученных моделей были усреднены для получения итогового значения метрик.
4. Общая выборка для дикторонезависимой модели (новостные тексты). Порождалась аналогично дикоторонезависимой выборке из п.2, однако в данном случае было 2 итерации (по количеству прочитавших текст дикторов).
5. Дикторозависимые выборки (художественные тексты). Для каждого диктора на материале его прочтения художественных повестей были составлены обучающая и тестовая выборки. Кросс-валидация не проводилась.
6. Дикторозависимые выборки (новостные тексты). Аналогично п. 4.
7. Меж-жанровые дикторозависимые выборки. Для дикторов, прочитавших новостные тексты, было составлено 2 пары выборок: в первой паре обучающий материал включал художественные тексты (как и в предыдущих случаях, 75% предложений), а тестовый – новостные (25% предложений), во второй паре – наоборот.

1.3. Анализ интонационной вариативности материала

Был проведен поверхностный анализ вариативности интонационных явлений, предсказание которых было целью данной работы. Это было сделано для примерной оценки сложности поставленных задач и базового уровня эффективности для системы предсказания.

Как правило, положение интонационного центра внутри синтагмы отличается слабой вариативностью: в большинстве случаев, центр расположен на последнем слове синтагмы (его ударном или, в случае моделей 07b и 11a, его заударном слоге). Центр может сдвигаться в вопросах (на вопросительное слово), если последнее слово не является знаменательным или для более выразительного прочтения и добавления интонационной выделенности. В художественных повестях корпуса интонационный центр был расположен на последнем слове синтагмы в 87% случаев; для новостных текстов это значение было выше – 92%. Таким образом, можно сделать вывод о том, что достаточно высокого уровня точности предсказания расположения интонационного центра можно достигнуть даже за счет простого привязывания центра к последнему слову синтагмы. При этом, в новостных текстах вариативность еще меньше, чем в художественных (и, следовательно, точность предсказания еще выше).

Вариативность интонационных моделей была проанализирована с точки зрения поверхностного анализа предложений: на простейших случаях (короткие односинтагменные предложения) была изучена зависимость между частотностью интонационных моделей и конечными знаками препинания в предложении (поскольку они в значительной степени отражают коммуникативный тип предложения, который, в свою очередь, является одним из главных факторов, влияющих на выбор интонационной модели). Для этого из художественных текстов были выбраны предложения длиной не менее 6 слов и без знаков препинания внутри; были учтены только те прочтения, где все предложение было реализовано как одна синтагма (90% случаев). Далее предложения были разделены на группы в зависимости от их конечного знака

препинания, и была подсчитана статистика употребимости интонационных моделей при их реализации.

Вопросительные предложения характеризовались максимальной среди найденных типов предложений согласованностью среди всех дикторов в выборе интонационной модели: 24% предложений были реализованы с одинаковой интонационной моделью всеми восемью дикторами; надо отметить, что данное значение согласованности является крайне низким, а то, что оно является максимальным среди найденных, позволяет получить представление о масштабе вариативности интонационного оформления в целом. Среди наиболее частотных моделей восклицательных предложений – 03 и 03a, 07, 08, 06c; исходя из данных моделей, можно предположить, что основными факторами, влияющими на выбор интонации вопросительного предложения, являются тип вопроса (специальный или общий), его длина (03 встречается реже для более длинных предложений, 03a – наоборот, для более коротких) и наличие эмфазы или эмоциональной окраски (например, недоумения). Дикторозависимые распределения моделей включали те же единицы, что показывает отсутствие четко выраженных дикторских предпочтений в выборе модели.

Следующим типом предложений по уровню согласованности дикторов являлись восклицательные предложения: в 14% случаев они были реализованы всеми дикторами с одинаковой интонацией. Наиболее употребительны среди них были модели 04, 02c, 02, также встречались 01a, 04a, 04b и окказионально – 05, 06, 06a. Таким образом, для данных предложений распространены не только специальные восклицательные интонационные модели (04, 05, 06, 06a), но и те, что в первую очередь ассоциируются с дополнительной выделенностью (02c, 02) или полной завершенностью (01a), либо используются в обращениях (04a) или просьбах (04b).

Для остальных типов предложений наблюдалось резкое падение уровня согласованности дикторов: для предложений, оканчивающихся точкой, в 5,8% случаев предложение было реализовано всеми дикторами одинаково, а для предложений с многоточием в конце данное значение было равно 0,9%. При

этом для первого типа наиболее употребительными моделями были: 01 и все подтипы, 02c, 02, 04b, 13; наблюдается значительное пересечение с восклицательными предложениями (02c, 02, 04b также частотны и для них), а также наличие интонации незавершенности (особенно часто – для коротких предложений). Для второй группы набор частотных моделей отличается большим разнообразием по сравнению с другими типами предложений: 01, 01a, 02, 02c, 04, 04b, 06, 13, 11, 11b – предложения с многоточиями могут реализовываться с интонацией завершенности, выделенности, вопроса, незавершенности или восклицания.

Данный анализ показывает, что даже при наличии значительных ограничений (небольшая длина предложения, односинтагменная реализация) предсказание интонационной модели исходя из поверхностного анализа предложения, не привлекающего лингвистических инструментов для морфологического, синтаксического и др. типов анализа, крайне затруднительно, учитывая слабую согласованность дикторов (максимальное ее значение составляет всего лишь 24%, минимальное – 0,9%) и высокую вариативность (почти для каждого типа предложений было возможно выделить 4 или более частотных модели). Это позволяет предположить, что использование более сложных подходов (в частности, нейросетевого) и задействование лингвистических инструментов оправдано, и изучение их эффективности важно и актуально для данной задачи.

2. Классификационные признаки

На выбор интонационной модели влияет большое число различных факторов, сложность формализации которых также широко варьируется: в то время как коммуникативный тип предложения можно в значительной степени определить по знаку препинания, более сложные факторы, такие, как возможность эмфатического выделения или уместность той или иной эмоциональной окраски, требуют анализа семантики предложения. В связи с этим, классификационные признаки, используемые в системе предсказания,

были разбиты на несколько основных групп, различающихся между собой по тому, с каким уровнем языка они взаимодействуют.

2.1. Пунктуационные признаки

Учет пунктуации необходим, поскольку помогает определить функциональное значение синтагмы (например, незавершенность, восклицание, вопрос и т. д.) и потому позволяет очертить множество наиболее вероятных интонационных моделей. Следующие пунктуационные признаки были учтены в системе предсказания:

1) **бинарный признак**, указывающий на то, **следовал ли** в тексте непосредственно за текущим словом какой-либо **знак пунктуации** (данный признак необходим, поскольку в нейронную сеть в качестве объектов для классификации подаются только слова, а знаки пунктуации удаляются);
2) **следующий за словом ближайший знак пунктуации** в текущем предложении; сочетание нескольких пунктуационных знаков (например, *'?!'*) считалось единым знаком пунктуации. Для данного признака было найдено 26 классов пунктуационных знаков.
3) **предшествующий слову ближайший знак пунктуации** в текущем предложении. Для данного признака было найдено 20 классов пунктуационных знаков (включая «нулевой» класс – отсутствие каких-либо знаков, предшествующих слову).
4) **бинарный признак**, указывающий, является ли предложение, к которому принадлежит данное слово, **последним предложением в абзаце**. Данный признак был учтен, поскольку, например, модель 01a, обозначающая полную завершенность, чаще всего встречается именно в конце абзаца.

2.1. Морфологические признаки

Предполагалось, что учет морфологических признаков важен в первую очередь для повышения точности предсказания расположения интонационного

центра, который может “притягиваться” к знаменательным частям речи, например, именам существительным или глаголам, если последнее слово в синтагме таковым не является. Следующие признаки входили в число морфлогических:

1) **частеречный тэг**. Морфологическая разметка проводилась с помощью модуля Deeppavlov (Burtsev et al., 2018) в терминах системы Universal Dependencies (Nivre et al., 2017). Количество морфологических тэгов, найденных в материале, было равно 16.
2) **бинарный признак**, указывающий, является ли словоформа **императивом**. Добавление признака обосновано наличием в системе Н. Б. Вольской интонационной модели 04b, которая зачастую употребляется при реализации императивных форм.

2.2. Лексические признаки.

Поскольку интонационное оформление в значительной степени зависит от семантики высказывания, его лексического наполнения и наличия эмоциональной окраски (которая также может быть частично отражена в лексике), при предсказании интонационных моделей может быть важно учитывать данные явления в какой-либо форме. Для отражения лексико-семантического содержания текста применялись **эмбеддинги BERT** (Devlin et al., 2018). Данная форма векторного представления слов превосходит более ранние аналоги по эффективности и успешно используется в большом числе разнообразных задач по естественной обработке языка (Xu et al., 2020; Annamoradnejad et al., 2020; Hayashi et al., 2019; Chen et al., 2019). Эмбеддинги BERT позволяют представлять значение слова в зависимости от его контекста: таким образом, эмбеддинги для одного и того же слова будут различаться, если оно употреблено в разных по содержанию предложениях. Обучение BERT происходит с применением трансформеров (Vaswani et al., 2017). Будучи мощным инструментом векторного семантического представления, эмбеддинги BERT зачастую заменяют собой все остальные лингвистические признаки,

включая синтаксические и морфологические. Поэтому в данном исследовании при изучении эффективности различных комбинаций признаков эмбеддинги использовались в качестве “базового” признака, к которому добавлялись все остальные.

В работе применялись предобученные эмбеддинги для русского языка из модели RuBERT (Kuratov et al., 2019), размерность которых составляла 768 значений.

2.3. Синтаксические признаки

Синтаксические признаки теоретически имеют большую значимость как для предсказания расположения интонационных центров, так и для определения интонационного оформления, поскольку просодическая структура в значительной мере связана с синтаксической и отражает ее в акустической стороне речи. Следующие признаки были включены в список синтаксических:

1) **синтаксический тэг**: класс синтаксического отношения, связывающего текущее слово и слово, от которого оно зависит. Синтаксический разбор текста осуществлялся с помощью модуля DeepPavlov, категории отношений были определены, как и в случае с морфологическими признаками, системой Universal Dependencies. Количество найденных в материале синтаксических тэгов было равно 36.
2) **относительный номер главного слова**: целое число, указывающее на расположение в предложении слова, от которого зависит текущее слово. Значение было равно 0, если слово являлось сказуемым, было положительным, если главное слово находилось в правом контексте текущего слова, и отрицательным, если оно находилось в левом контексте.

2.4. Фонетические признаки

Единственным фонетическим признаком был бинарный признак, указывающий на наличие или отсутствие **синтагматической границы** после

текущего слова. Предсказание синтагматических границ является важной частью предсказания интонационного оформления для синтеза речи. Эксперименты по решению данной задачи были также проведены автором настоящей работы на материале корпуса CORPRES (Menshikova & Kocharov, 2019); было показано, что предсказание синтагматических границ на основе текстовых признаков может быть осуществлено для русского языка с высокой точностью — до 90.4% F1-меры при использовании нейросетевого подхода. Однако в текущем исследовании предсказание синтагматических границ не проводилось, и их расположение извлекалось напрямую из просодической аннотации для речи дикторов. Совмещение двух предсказательных моделей — для синтагматических границ и для интонационных моделей — остается делом будущих работ.

Все признаки были конвертированы в вектора при помощи унитарного кода. Далее признаки были конкатерированы в единый вектор размерностью в 876 значений.

2.5. Целевые классы

Целью метода было предсказание типа интонационной модели, без учета подтипов, выделяемых в системе Н. Б. Вольской: таким образом, модели 01, 01a, 01b и т.п. считались единым классом. Конвертация классов из системы Н. Б. Вольской в классы системы Е. А. Брызгуновой для оценки эффективности каждой системы описания осуществлялась по правилам, установленным в Таблице 2.

Таблица 2. Соответствия между ИК системы Е. А. Брызгуновой и моделями Н. Б. Вольской, принятые в данной работе.

Модель по системе Е. А. Брызгуновой	Модели по системе Н. Б. Вольской
1	01, 09, 10
2	02, 03, 04

3	07, 11
4	08, 13
5	05
6	06, 12
7	—

Класс слова был представлен в нейронной сети с помощью унитарного кода. Также был добавлен нулевой класс — отсутствие интонационного центра. Таким образом, слова, классифицированные как принадлежащие к нулевому классу, не считались носителями интонационного центра, в то время как слова, отнесенные предсказательной моделью к любому другому классу, считались центром синтагмы.

В итоге, **класс** для каждого слова кодировался с помощью 7-размерного или 14-размерного вектора, в зависимости от системы интонационного описания, используемой для представления интонационных моделей (7 значений для системы Е. А. Брызгуновой, 14 значений для системы Н. Б. Вольской).

3. Метод

3.1 Метрики

В связи с тем, что предсказание проводилось для двух аспектов интонационного оформления — расположения интонационного центра и типа интонационной модели, для оценки работы системы было введено две категории параметров.

В первую категорию вошли метрики точности (Precision, Pr), полноты (Recall, Re) и F1-меры (F1-measure, F1), традиционно используемые в машинном обучении для оценивания систем, работающих в условиях несбалансированности классов и нацеленных на предсказание редких явлений. Поскольку на одну синтагму, в среднем состоящую из 4-5 слов, приходится

лишь один интонационный центр, слова-носители интонационных центров можно считать малочисленным классом по отношению ко всем остальным словам. Метрика точности показывает, какая доля слов, для которых модель предсказала класс носителя ядра синтагмы, в действительности (в дикторской реализации) являлась таковыми:

$$Precision = \frac{TP}{TP + FP},$$

где TP — количество верно определенных интонационных центров, FP — количество слов, не являющихся носителями ядра, но определенных нейросетью как таковые.

Метрика полноты показывает, для какой доли слов, в действительности являющихся носителями интонационного центра, модель предсказала наличие интонационного центра:

$$Recall = \frac{TP}{TP + FN},$$

где TP — количество верно определенных интонационных центров, FN — количество слов, являвшихся при дикторском прочтении носителями ядра, но не имевших данный класс при нейросетевом предсказании.

F1-мера является средним гармоническим полноты и точности и показывает сбалансированную оценку эффективности системы с точки зрения этих метрик:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall},$$

Вторая категория метрик — оценка эффективности предсказания интонационных моделей — оценивалась с помощью точности (Accuracy, Acc) по формуле:

$$Accuracy = \frac{T}{T + F},$$

где T — число синтагм с верно предсказанными интонационными моделями, а F — число синтагм с неверно предсказанными интонационными моделями. При этом, при вычислении данной метрики проводилась

постобработка результатов: если в синтагме не было предсказано ни одного интонационного центра, предсказанной моделью считалась та, что имела наибольшую вероятность для последнего слова синтагмы.

Метрика точности предсказания, однако, имеет недостаток, заключающийся в том, что на ее значение крайне слабо влияют низкочастотные интонационные модели. Таким образом, данная метрика не позволяет отличить предсказательные модели, допускающие значительную степень вариативности в интонационном оформлении, от моделей, всегда выбирающих высокочастотные альтернативные варианты в местах, где могли бы быть употреблены редкие интонационные модели. Данный аспект оценки, однако, можно считать важным, поскольку качественный метод предсказания должен учитывать и отражать в своих результатах вариативность и разнообразие интонационного оформления. В связи с этим для оценки эффективности предсказания также использовалась метрика соответствия распределения моделей (далее CPM):

$$MDM = 1 - \frac{1}{N}\sum_{i=0}^{N}\frac{|actual_i - predicted_i|}{actual_i + predicted_i},$$

где N — количество типов интонационных моделей (13 или 6 в зависимости от используемой системы интонационного описания), i — текущий тип интонационной модели, $actual_i$ — число раз, когда дикторы употребили интонационную модель данного типа, $predicted_i$ — число раз, когда нейросетевая модель предсказала интонационную модель данного типа. Чем меньше разница между числом предсказанных и числом реализованных дикторами моделей каждого типа, тем больше значение метрики CPM. Таким образом, она показывает, насколько точно распределение предсказанных интонационных моделей соответствует распределению реализованных дикторами моделей; при этом все интонационные модели, вне зависимости от частотности, имеют равный вес, поэтому если метод предсказывает только высокочастотные модели, игнорируя возможность появления низкочастотных, это заметно и негативно скажется на значении метрики CPM.

Значения всех метрик варьировались в диапазоне между 0 и 1; для удобства представления значения также домножались на 100.

3.2 Описание архитектуры и параметров

В качестве нейросетевой архитектуры для решения задачи предсказания интонационного оформления была выбрана двунаправленная нейронная сеть долгой краткосрочной памяти (Bidirectional Long Short-Term Memory, BiLSTM). Данная архитектура эффективно используется при моделировании последовательностей: к таковым относится значительная часть языковых задач, как, например, частеречная разметка (Liu et al., 2018) или предсказание синтагматических границ (Bach et al., 2019). Как и LSTM, BiLSTM принадлежит к рекуррентным нейронным сетям, позволяющим учитывать контекст классифицируемого объекта, что чрезвычайно важно для предсказания интонационного оформления, где контекст слова-носителя интонационного центра в большей степени определяет контур, чем характеристики самого слова. BiLSTM, состоящая из двух LSTM-компонентов, позволяет моделировать последовательность исходя из признаков правого и левого контекста объекта (в отличие от LSTM слоев, которые учитывают только левый контекст).

Нейросетевая модель была построена при помощи модулей Keras (Chollet, 2015) и Tensorflow (Abadi et al., 2015). Параметры обучения модели были определены эвристически с помощью выборки для тестирования параметров модели, описанной в разделе 1.3: модели с разными значениями параметров были обучены и протестированы на данной выборке; в итоговой архитектуре были использованы параметры, для которых наблюдались наилучшие значения метрик. С помощью данного подхода были в указанном порядке выбраны значения следующих параметров: количество слоев и их (BiLSTM или LSTM), количество нейронов в каждом слое, размер бэтча (число объектов обучающей выборки, которые модель обрабатывает перед обновлением значений весов нейронов). Результаты для различных значений данных параметров представлены в Таблицах 3, 4, 5 соответственно.

Таблица 3. Сравнение эффективности моделей с разным количеством и типом слоев.

Тип слоя	Количество слоев	Precision	Recall	F1	Accuracy	CPM
LSTM	1	91,4	89,9	90,7	**64,7**	61,2
	2	93,1	88,2	90,6	64,3	55,1
	3	92,7	89,3	90,9	64,3	54,1
BiLSTM	1	**93,2**	89,5	91,3	64,1	67,5
	2	92,9	89,9	**91,4**	64,5	**67,8**
	3	92,8	**90,1**	**91,4**	**64,7**	60,9

Результаты, представленные в Таблице 3, показывают, что использование BiLSTM слоев систематически приводит к более качественным результатам с точки зрения метрики CPM и предсказания расположения центра. Значение CPM, в отличие от предсказания расположения центра, также варьируется и с изменением количества слоев, при этом преимущественно падает с увеличением их количества. Предположительно, это можно объяснить тем, что более сложные модели — в частности, имеющие большее число слоев — требуют больше обучающего материала, а его нехватка приводит к переобучению, из-за чего модель “пренебрегает” редкими интонационными конструкциями и предсказывает вместо них наиболее частотные.

Метрика точности предсказания интонационной модели же с изменением количества и типа слоев практически не меняется.

Было принято решение использовать в итоговой архитектуре два BiLSTM слоя, поскольку эта комбинация одновременно имела сравнительно высокое значение F1-меры и CPM.

Таблица 4. Сравнение эффективности моделей с разным количеством нейронов в каждом слое.

Количество нейронов	Precision	Recall	F1	Accuracy	CPM
64	92,9	90,5	91,7	**65,7**	58,8
128	93,3	90,4	91,9	64,8	66,1
256	**93,6**	90,0	91,8	63,8	71,3
512	93,1	**90,9**	**92,0**	63,5	**75,2**

Из Таблицы 4 можно увидеть, что увеличение количества нейронов в каждом слое, в отличие от увеличения числа слоев, приводит к улучшению значения CPM и F1-меры. Несмотря на то, что значение точности предсказания интонационных моделей при этом падает, было принято решение в пользу более сложной архитектуры, и в итоговой модели было использовано 512 нейронов в каждом слое.

Таблица 5. Сравнение эффективности моделей с разным размером бэтча.

Значение бэтча	Precision	Recall	F1	Accuracy	CPM
32	93,7	89,2	91,4	64,4	**73,1**
64	**94,1**	88,8	91,3	63,9	64,9
128	93,2	89,9	91,6	64,4	62,5
256	92,7	90,8	91,8	63,9	59,8
512	92,3	90,7	91,5	**64,7**	54,8
1024	91,7	**92,2**	**91,9**	64,5	58,0

Известно, что меньший размер бэтча замедляет скорость обучения нейросети, однако улучшает ее результаты (Kandel & Castelli, 2020). Предполагается, что меньший размер бэтча способствует появлению дополнительного шума в обучающей выборке, что позволяет достичь лучшей обобщающей способности нейросети. В данной работе данный эффект размера бэтча проявляется скорее на примере CPM и точности предсказания

расположения центра, но не на полноте и точности предсказания типа модели. Тем не менее, был выбран размер бэтча 64, позволяющей достичь лучшей точности предсказания расположения центра, а также являющийся промежуточным между 32 (демонстрирующим наилучшее значение CPM) и более крупными значениями, приводящими к улучшению показателей остальных метрик.

Таким образом, итоговая архитектура состояла из двух BiLSTM слоев с 512 нейронами в каждом и коэффициентом drop-out в 0.3, одним слоем Dense с 20 нейронами и функцией активации ReLU[1], одним слоем Drop-out[2] с коэффициентом 0.3, и одним слоем Dense с функцией активации softmax. В качестве функции потерь была выбрана категориальная кросс-энтропия, в качестве оптимизатора — адаптивная оценка момента (Adam[3]). Количество эпох определялось динамически: обучение модели заканчивалось, когда на протяжении 5 эпох не было улучшения в результатах функции потерь, однако, для количества эпох также было установлено максимальное количество в 100. Размер бэтча был равен 64.

Максимальная длина передаваемой в нейросеть последовательности была равна 25 словам (значение было выбрано эмпирически на основе графиков с частотностью предложений различной длины в материале исследования). В случае, если длина предложения составляла менее 25 слов, в массив предложения добавлялось нужное число нулевых векторов.

---

[1] Agarap A. F. Deep learning using rectified linear units (relu) //arXiv preprint arXiv:1803.08375. – 2018.

[2] Cicuttin A. et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting //2016 Int. Conf. Adv. Electr. Electron. Syst. Eng. ICAEES 2016. – 2017. – Т. 15. – С. 520-525.

[3] Kingma D. P., Adam B. J. a method for stochastic optimization. 2014 //arXiv preprint arXiv:1412.6980. – 2018. – Т. 9.

4. Результаты

4.1 Сравнение комбинаций признаков

Для определения наилучшей комбинации признаков для обучения системы предсказания были проведены специальные эксперименты: было обучено несколько моделей, использующих различные признаки, и их результаты были сравнены для выбора оптимального сочетания признаков для дальнейших экспериментов. Как было упомянуто ранее в разделе 2.2, эмбеддинги BERT были выбраны в качестве базового признака, к которому постепенно добавлялись остальные (в порядке возрастания сложности извлечения признака). Однако также был проведен эксперимент без использования эмбеддингов. Обучение проводилось на общей выборке художественных текстов для дикторонезависимой модели, тестирование — на валидационной части этой выборки (20% предложений изымались из обучающей выборки и использовались для тестирования). Валидационная выборка использовалась по той причине, что отбор признаков, фактически, все еще является подбором параметров модели, и использование тестовой выборки для этой цели было бы некорректно. Результаты приведены в Таблице 6.

Таблица 6. Сравнение эффективности комбинаций признаков.

Набор признаков	Precision	Recall	F1	Accuracy	CPM
BERT	83,2	68,8	75,3	56,9	59,9
BERT, пунктуация	84,2	71,2	77,1	58,4	58,1
BERT, пунктуация, морфология	84,0	73,4	78,3	57,8	55,1
BERT, пунктуация, морфология, синтаксис	84,3	74,5	79,1	59,6	60,2
BERT, пунктуация, морфология, синтаксис, фонетика	**90,6**	**88,2**	**89,4**	**60,2**	64,5
Пунктуация, морфология, синтаксис,	89,7	85,1	87,4	59,8	**64,7**

фонетика					

Использование всех групп признаков позволяет достичь наилучших значений всех метрик, кроме CPM (значение данной метрики в этом случае всего на 0,2% меньше, чем максимальное из полученных); в связи с этим, во всех остальных экспериментах были задействованы все извлекаемые признаки. Однако исключение эмбеддингов BERT снижает показатели F1-меры и точности предсказания интонационной модели на 2% и 0,4% соответственно, а также улучшает CPM, из чего можно предположить, что эмбеддинги BERT не являются самым важным и эффективным признаком — сравнительный вклад других признаков зачастую гораздо больше. Вероятно, это связано с недостаточно большой по размеру обучающей выборкой, из-за чего обучение модели с высокоразмерными эмбеддингами затруднено, и их потенциал не раскрывается полностью.

Анализ вклада каждой группы признаков показывает, во-первых, что добавление пунктуационных признаков повышает общее качество предсказания расположения интонационного центра (F1-мера стала выше на 1,8%), а также точность предсказания интонационной модели (улучшение на 1,5%), однако несколько снижает значение CPM. Можно предположить, что это происходит потому, что пунктуационные признаки в большинстве случаев не позволяют разграничить взаимозаменяемые высоко- и низкочастотные интонационные модели, поскольку те встречаются в предложениях одних и тех же коммуникативных типов: например, модели 10, 11, 12 и 13 являются интонациями незавершенности и преимущественно взаимозаменяемы, однако 11 является самой частотной из всех. Во-вторых, морфологические признаки повышают полноту и потому F1-меру (на 2,2% и 1,2%), однако негативно сказываются на качестве предсказания интонационного оформления с точки зрения обеих метрик; потому морфологические признаки могут быть рекомендованы для предсказания именно расположения интонационного центра, но не контура. Заметен вклад синтаксических признаков во все аспекты

предсказания: 0,8% для F1-меры, 1,8% для точности предсказания интонационного оформления, 5,1% для CPM. Это подтверждает на практике идею о том, что учет синтаксической структуры важен и полезен для предсказания интонационного оформления, вероятно, даже больше, чем большинство других признаков. Фонетические признаки также входят в ряд наиболее значимых, добавляя 10,3%, 0,6%, 4,3% соответственно к значениям метрик F1-меры, точности предсказания контура и CPM (повышение CPM связано, вероятно, с тем, что интонационное оформление в некоторых случаях зависит от длины синтагмы, и потому на предсказание оформления влияют конкретные синтагматические границы). Тем не менее, важно помнить, что при использовании предсказанных синтагматических границ вместо реально реализованных диктором вклад фонетических признаков снизится.

4.2 Сравнение интонационно сбалансированной и случайной выборок

Итоговая архитектура, для которой были экспериментально выбраны значения параметров и набор признаков, была также дополнительно обучена и протестирована на интонационно сбалансированной выборке (п. 2 раздела 1.2) и на выборке со случайным разделением на обучающий и тестовый наборы предложений (п. 1 раздела 1.2). Сравнение интонационно сбалансированной выборки именно с данной случайной выборкой проводилось потому, что для них обеих, в отличие от прочих дикторонезависимых выборок, не проводилась кросс-валидация и учет дикторов и предложений при составлении тестового и обучающего набора; таким образом, учет реализованных в предложении интонационных моделей — единственный фактор, различающий данные выборки. Целью эксперимента было установить, полезно ли учитывать интонационные модели, встретившиеся в предложении, при составлении обучающей выборки, и нужно ли проводить эту процедуру в дальнейших экспериментах.

Таблица 7. Сравнение интонационно сбаланированной выборки и выборки со случайным разбением на обучающий и тестовый наборы.

Тип выборки	Precision	Recall	F1	Accuracy	CPM
Интонационно сбалансированная	93,4	90,4	91,9	64,9	82,5
Случайная выборка	**93,6**	**90,6**	**92,0**	**66,0**	**82,6**

Из таблицы можно увидеть, что эффективность модели для интонационно сбалансированной выборки ниже с точки зрения всех метрик. Поскольку балансирование обучающей выборки в интонационном аспекте не приносит улучшения, в дальнейшем данная процедура не проводилась, и разбиение на обучающую и тестовую выборки было случайным.

4.3 Результаты для дикторонезависимых моделей на художественных текстах

Две модели было обучено и протестировано на дикторонезависимой выборке из художественных текстов (проводилась кросс-валидация по диктору, предложения из тестовой выборки не встречались в обучающей). В первой модели для описания целевых классов использовалась система интонационного описания Н. Б. Вольской, во второй модели — система Е. А. Брызгуновой. Результаты представлены в Таблице 8.

Таблица 8. Сравнение дикторонезависимых моделей, кодирующих целевые классы с помощью различных систем интонационного описания.

Система описания	Precision	Recall	F1	Accuracy	CPM
Н. Б. Вольской	**89,9**	**87,5**	**88,7**	**61,2**	**65,2**
Е.А. Брызгуновой	89,8	87,1	88,4	59,2	59,9

Во-первых, значения в таблице показывают, что использование системы описания Н. Б. Вольской позволяет достичь лучших значений всех пяти метрик. Особенно значителен разрыв для точности предсказания модели (2%) и CPM (5,3%). Это позволяет предположить, что данная система предпочтительна при решении задачи предсказания интонационного оформления: не только потому, что получает достичь более качественных результатов предсказания, но и потому, что предоставляет больше информации о предложении и об интонационном контуре для последующих этапов синтеза речи. В связи с этим, во всех последующих экспериментах представлены только результаты для моделей, использующих систему Н. Б. Вольской.

Во-вторых, стоит отметить значения метрик, полученных в данном эксперименте, поскольку они являются итоговыми для дикторонезависимой модели на материале художественных текстов: 88,7% F1-меры, 61,2% точности предсказания модели, 65,2% CPM.

4.4 Результаты для дикторозависимых моделей на художественных текстах

Результаты для дикторозависимых моделей, обученных на материале художественных текстов, описаны в Таблице 9. Кодовое название каждого диктора содержит в себе указание на пол диктора (M — мужчина, F — женщина) и порядковый номер.

Таблица 9. Значения метрик для дикторозависимых моделей и статистические значения для каждой метрики.

Диктор	Precision	Recall	F1	Accuracy	CPM
M1	**90,1**	**87,0**	**88,6**	55,2	78,1
M2	87,5	85,8	86,7	**66,1**	74,8
M3	88,7	85,7	87,2	63,4	69,7
M4	86,7	82,4	84,5	61,3	65,0

F1	88,9	85,3	87,0	54,4	76,9
F2	90,0	85,7	87,8	62,9	**78,5**
F3	87,4	86,9	87,1	61,9	65,7
F4	84,8	83,6	84,2	62,6	73,7
Среднее	88,0	85,3	86,6	60,9	72,8
Стандартное отклонение	1,8	1,6	1,5	4,1	5,3
Максимальное	90,1	87,0	88,6	66,1	78,5
Минимальное	84,8	82,4	84,2	54,4	65,0

Из данных таблицы можно сделать общий вывод о том, что качество дикторозависимых систем хуже, чем дикторонезависимой, с точки зрения определения расположения интонационного центра: в то время как дикторонезависимая система показывает значение F1-меры в 88,7% (усредненное по всем кросс-валидационным итерациям), для дикторозависимых систем максимальным значением является 88,6% (для диктора M1), а минимальным — 84,2%. Происходит это в первую очередь именно за счет снижения полноты в дикторонезависимых моделях по сравнению с общей. С другой стороны, семь из восьми дикторозависимых моделей показывают более высокие значения CPM, чем дикторозависимая, а шесть из восьми — и более высокие значения точности предсказания интонационной модели. Среднее значение CPM для дикторозависимых моделей равно 72,8%, в то время как для общей модели данная метрика равна 65,2% соответственно. Исходя из этих данных, можно предположить, что разделение модели предсказания интонационного оформления на два компонента, обучаемых независимо друг от друга — предсказание расположения центра и предсказание интонационной модели — могло бы стать оптимальным решением, особенно, если обучать первый компонент на дикторонезависимом материале, а второй — для каждого диктора индвидуально.

Еще одной основной характеристикой дикторозависимых моделей является большой разброс в показателях их эффективности: стандартное отклонение для точности определения интонационной модели составляет 4,1%, а разница между наилучшей (диктор M2) и худшей моделью (диктор F1) — 11,7%. Анализ распределений интонационных моделей в тестовых выборках этих двух моделей показал, что диктор F1 отличается большей вариативностью в выборе интонационного оформления; например, у данного диктора менее ярко, чем у диктора M2, выражены предпочтения среди интонаций незавершенности. В Таблице 10 можно увидеть, что употребление модели 11, наиболее частотной интонации незавершенности, составляет у диктора M1 69% от общего числа интонационных моделей со значением незавершенности. У диктора F1, однако, то же значение составляет лишь 37,5%; остальные модели — 10, 12, 13 — у диктора F1 более частотны, чем у диктора M2. Это показывает, что речь диктора F2 более интонационно разнообразна и, следовательно, несколько менее предсказуема; из-за этого качество предсказательной модели ниже, чем для диктора M2, имеющего более четко выраженные предпочтения и значительно реже использующего менее частотные конструкции. С другой стороны, значение CPM, демонстрируемое моделью для диктора F1, довольно высоко (76,9%) и превосходит то же значение для модели диктора M2 (74,8%). Схожая ситуация наблюдается для дикторозависимой модели M1, также имеющей одно из самых низких значений точности предсказания оформления (55,2%) и одно из самых высоких значений CPM (78,1%). Можно предположить, что хотя интонационное оформление для дикторов, демонстрирующих большее разнообразие, хуже поддается предсказанию, предсказательные модели, обученные на таком материале, все же гораздо чаще предсказывают низкочастотные интонационные конструкции (подтверждение можно увидеть также в Таблице 10: последние два столбца показывают, что для диктора F1 модели 10, 12, 13 встречались чаще, чем для диктора M2); из-за этого показатель CPM довольно высок, несмотря на низкое значение точности предсказания оформления (по сравнению с другими дикторами). В целом,

однако, однозначной корреляции между CPM и точностью предсказания интонационного оформления не наблюдается.

Таблица 10. Сравнение распределения интонационных моделей, обозначающих незавершенность, для диктора F1 (модель данного диктора показала худшее значение точности предсказания интонационного оформления) и диктора M2 (лучшее значение точности). Для обоих дикторов встретилось порядка 1200 синтагм, оформленных с помощью интонационных моделей незавершенности.

Модель	Доля модели от всех дикторских употреблений интонаций незавершенности в тестовой выборке, %		Доля модели от всех предсказанных дикторозависимой моделью интонаций незавершенности, %	
	Диктор F1	Диктор M2	Диктор F1	Диктор M2
10	24,5	14,5	29,6	14,5
11	37,5	69	43,4	79,1
12	27,4	9,7	21,5	4,5
13	10,6	6,8	5,5	1,9

4.5 Результаты для моделей, обученных на материале новостных текстов

Эксперименты, результаты которых представлены в данном разделе, были проведены в целях сравнения эффективности предсказательной модели на материале художественных и новостных текстов. Значения метрик для дикторонезависимых моделей, обученных и протестированных на новостных текстах, представлены в Таблице 11. Как можно было предположить, в случае с обоими дикторами, для которых были доступны записи чтения новостных текстов, показатели эффективности моделей для новостных текстов выше, чем аналогичные показатели для художественных текстов.

Исключением является только значение CPM для диктора M1: в модели, обученной для художественных текстов, оно выше; этот факт можно связать с

одним из недостатков метрики CPM: ее значение подвержено искажению в случае, если какая-либо модель слишком редко используется. В новостных текстах для диктора M1 это случилось с моделями 04, 06 и 07. В тестовой выборке для новостных текстов диктор M1 употребил модель 04 один раз, модель 06 — также один раз, а модель 07— два раза; поскольку нейросетевая модель ни разу не предсказала данных интонационных конструкций, значение CPM для классов данных моделей оказалось равно нулю, и потому среднее CPM модели значительно снизилось. Если исключить эти интонационные конструкции из рассмотрения, CPM для модели диктора M1 на новостных текстах оказывается равно 80,6%, превосходя аналогичное значение для художественных текстов.

Таблица 11. Результаты для дикторозависимых моделей на материале новостных текстов (для сравнения приведены результаты для аналогичных дикторозависимых моделей на художественных текстах).

Диктор	Выборка	Precision	Recall	F1	Accuracy	CPM
M1	Новост.	**92,3**	**94,2**	**93,2**	**60,5**	58,6
	Худож.	90,1	87,0	88,6	55,2	**78,1**
F3	Новост.	**91,0**	**89,4**	**90,2**	**67,1**	**72,0**
	Худож.	87,4	86,9	87,1	61,9	65,7

В Таблице 12 представлены результаты для дикторонезависимой модели, обученной на новостных текстах. Ее эффективность также выше, чем эффективность модели для художественных текстов, даже несмотря на то, что их сравнение не вполне корректно: модель для новостных текстов была обучена на материале чтения одного диктора (и протестирована на материале чтения второго), в то время как модель для художественных текстов была обучена на значительно более репрезентативном материале (материал чтения семи дикторов).

Таблица 12. Результаты для дикторонезависимой модели на материале новостных текстов (для сравнения приведены результаты для дикторонезависимой модели на художественных текстах; показатели для нее подсчитаны на основе двух кросс-валидационных итераций, где материал, порожденный дикторами M1 и F3, составлял тестовую выборку).

Выборка	Precision	Recall	F1	Accuracy	CPM
Новостн.	**91,6**	**90,8**	**91,2**	**57,9**	**64,0**
Худож.	89,6	88,8	89,2	56,6	61,3

Таким образом, новостные тексты, являясь менее интонационно разнообразными и эмоционально насыщенными, несколько лучше поддаются предсказанию, как в плане расположения интонационного центра (который реже подвергается смещению, чем в художественных текстах), так и в плане предсказания интонационных моделей. Величина разницы в качестве, однако, зависит от конкретного диктора: например, для диктора M1 разница в значении F1-меры между новостной и художественной моделью более значительна, чем для диктора F3.

4.6 Результаты для меж-жанровых выборок

Для изучения эффективности модели в условиях значительно различающихся по стилю обучающей и тестовой выборок были проведены эксперименты с использованием меж-жанрового материала: обучение модели проводилось на текстах одного стиля, а тестирование — на текстах другого. Значения метрик представлены в Таблице 13. Представленные модели были дикторозависимыми. Стоит в первую очередь отметить значительное снижение уровня качества предсказания интонационных моделей — для всех дикторов и комбинаций выборок; значение CPM для диктора F3 в данном эксперименте достигло рекордно низких значений — 38,1% и 48,1% (но и в то же время, для диктора M1 CPM выше при использовании художественных текстов в качестве обучающего материала и новостных в качестве тестового, чем для чисто

новостной модели для данного диктора). Тенденция также очевидна для качества предсказания расположения интонационного центра: F1-мера во всех случаях ниже, чем при использовании однородных обучающих и тестовых выборок. Тем не менее, снижение F1-меры в условиях разнородных выборок менее радикально, чем для метрик CPM и точности предсказания контура, а в некоторых условиях и вовсе довольно мало (0,3% снижения для диктора M1 при обучении на новостных текстах и тестировании на художественных по сравнению с обучением и тестированием на художественных). Несмотря на это, можно сказать, что для предсказания интонационных моделей предпочтительней все же использовать художественные тексты в качестве обучающей выборки: в этом случае падение качества менее значительное, чем при использовании новостных текстов.

Таблица 13. Результаты меж-жанрового обучения и тестирования дикторозависимых моделей.

Диктор	Обучающая выборка	Тестовая выборка	Precision	Recall	F1	Accuracy	CPM
M1	Новостн.	Худож.	87,5	**89,1**	88,3	41,1	50,9
	Худож.	Новостн.	**94,0**	88,3	**91,1**	**51,9**	**65,4**
F3	Новостн.	Худож.	86,0	83,6	84,8	45,1	**48,1**
	Худож.	Новостн.	**89,8**	**87,1**	**88,4**	**56,7**	38,1

4.7 Анализ ошибок предсказательной модели

4.7.1 Анализ матрицы смешения

На рис. 1 изображена матрица смешения для дикторонезависимой модели, обученной на художественных текстах и предсказывающей интонационные модели в терминах системы Н. Б. Вольской. По вертикали указаны предсказанные классы, по горизонтали — реализованные дикторами. Анализ матрицы показывает, что наиболее верно система предсказывает нейтральную

интонацию завершенности (модель 01), общих вопросов (модель 07) и примечаний и комментариев (модель 09); близки к ним по точности определения нейросетью интонация эмоционального вопроса (модель 04), интонация специального вопроса (модель 03) и интонация незавершенности, реализуемая с помощью модели 11. К наиболее распространенным ошибкам, исходя из матрицы, можно отнести смешение различных интонаций незавершенности, к которым относятся модели 10-13, а также модель 02, которая может употребляться в тех же контекстах, что и интонация незавершенности — например, при оформлении однородных членов предложения (один член ряда может быть выделен с помощью интонации незавершенности, чтобы показать, что за ним будет следовать продолжение, либо с помощью модели 02 для того, чтобы противопоставить его следующим членам, “выделить” его). Другой общей ошибкой является смешение эмфатических и нейтральных интонаций, имеющих одно коммуникативное значение (модели 07 и 08; синтагмы, реализованные дикторами эмоционально окрашенные модели 08, зачастую ошибочно предсказываются как нейтральные представители модели 07). Также зачастую на месте дикторской модели 04 (восклицание, обращение или просьба) предсказываются модели 01 и 02 (нейтральная завершенность, завершенность с выделением), однако, поскольку все эти интонационные модели имеют нисходящий контур в интонационном центре, это можно считать менее критичной ошибкой. Весомой ошибкой можно считать, однако, смешение моделей 06 и 07 (на месте обеих в дикторском прочтении обычно 06), поскольку они различаются по своей акустической манифестации, а также могут передавать различные значения (модель 07 — исключительно вопросительная интонация, модель 06 может помимо интонации вопроса также передавать восклицание). Тем не менее, ситуация смешения данных моделей происходила редко: модель 06 низкочастотна (это же объясняет тот факт, что она заменяется высокочастотной моделью 07).

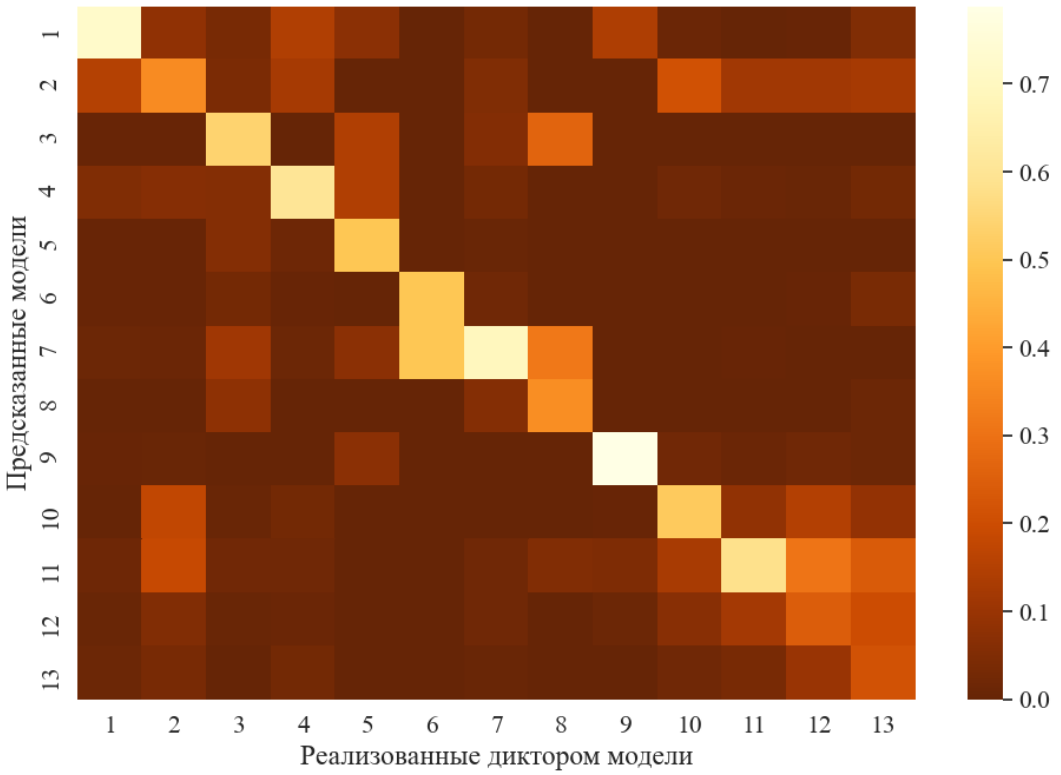


Рис. 1. Матрица смешения между интонационными моделями, предсказанными системой (по вертикали), и интонационными моделями, предсказанными дикторами тестовой выборки (по горизонтали).

Чтобы понять, насколько критичными являются ошибки системы, можно сравнить матрицу с Рисунка 1 с матрицей, показанной на Рисунке 2: матрицей смешения для одного из дикторов корпуса (M3), где сопоставлены его выборы интонационного оформления и выборы других дикторов для тех же слов.

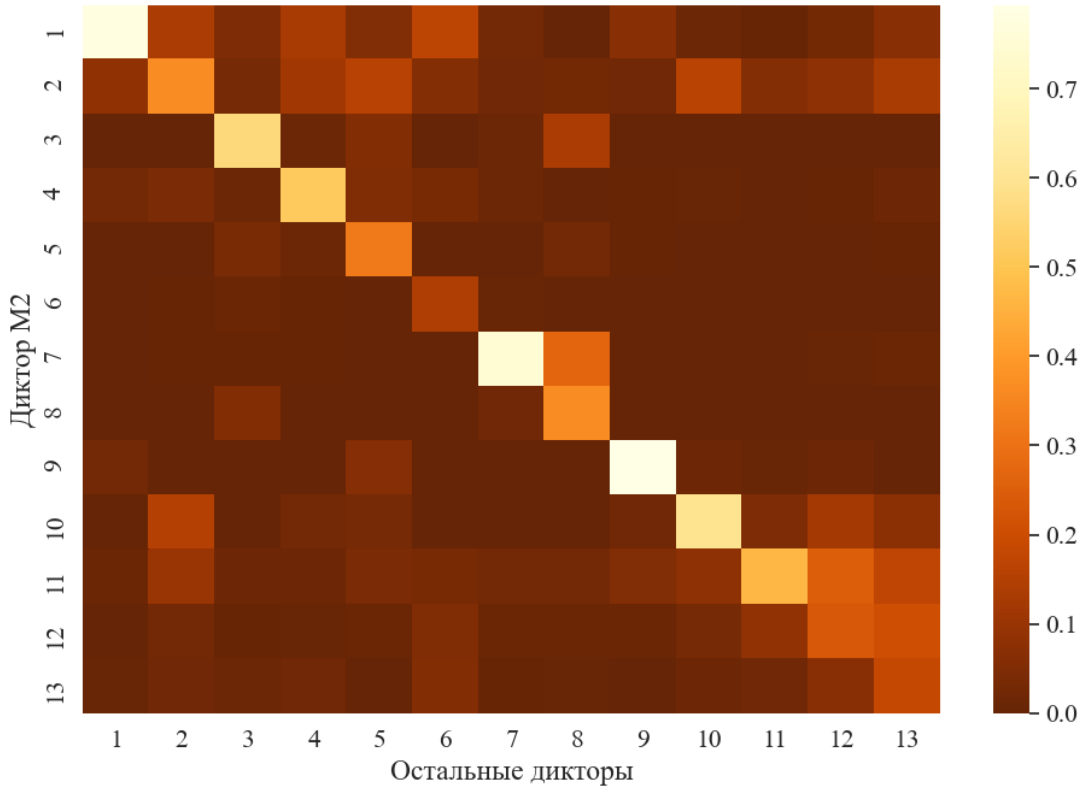


Рис. 2. Матрица смешения для диктора M3. Представлено сравнение реализованных диктором моделей (по вертикали) и моделей, реализованных остальными дикторами корпуса на тех же словах (по горизонтали).

По вертикали указаны модели, реализованные данным диктором, по горизонтали указаны модели, реализованные остальными дикторами корпуса на тех же словах. При сравнении Рисунков 1 и 2 можно увидеть, что они в значительной мере напоминают друг друга: на Рисунке 2 также наблюдается смешение интонационных моделей незавершенности (вместе с моделью 02), моделей 04, 05, 06 с моделями 01, 02 (эмфатическое или эмоционально окрашенное прочтение вместо более нейтрального), моделей 08 и 03 (разные акустические манифестации вопросов), 08 и 07. Матрицу смешения для модели предсказания отличает ошибка, касающаяся интонационных моделей 06 и 07. Однако смешение данных классов, хотя и в меньшей степени, присутствует на матрице для другого диктора корпуса (Рисунок 3). Это позволяет предположить, что наиболее частотные ошибки зачастую предсказательной модели не имеют

критичного характера, а являются адекватными альтернативными вариантами интонационного оформления синтагмы.

Матрицы смешения для всех дикторов корпуса можно найти в Приложении 1.

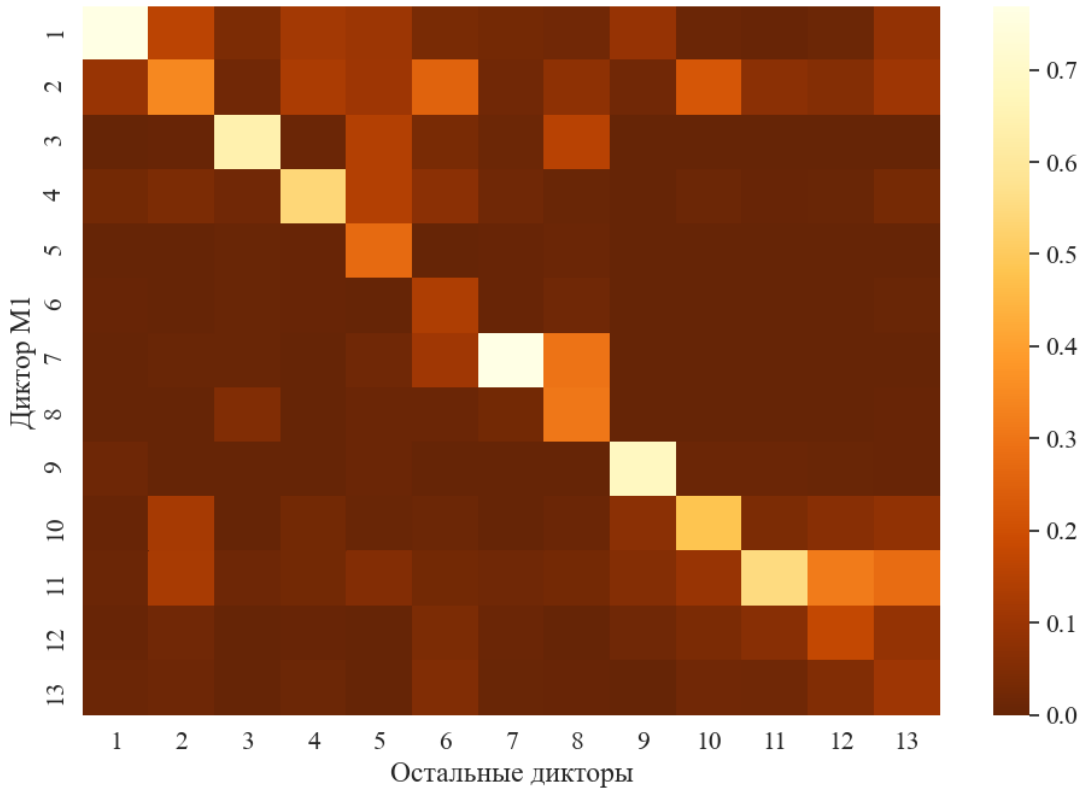


Рис. 3. Матрица смешения для диктора M1. Представлено сравнение реализованных диктором моделей (по вертикали) и моделей, реализованных остальными дикторами корпуса на тех же словах (по горизонтали).

Интересной также может показаться матрица смешения для меж-жанровой дикторонезависимой модели, обученной на художественных текстах и протестированной на новостных (описана в разделе 4.6). В данных условиях система предсказания показала один из худших результатов с точки зрения точности предсказания интонационных моделей. На матрице смешения можно увидеть, что наиболее надежно модель предсказывает интонацию общего вопроса (модель 07), а наиболее точно — интонацию завершенности (модель 01). Однако интонация общего вопроса (модель 07) также часто

появляется на месте дикторской интонации специального вопроса (модель 03), что является довольно весомой ошибкой; туда же можно отнести предсказание модели 01 (нейтральной завершенности) и моделей незавершенности на месте интонации восклицания (модель 04). На месте другой интонации восклицания — 05 — всегда появляется модель 03. В целом, наблюдаемое смешение интонационных моделей подтверждает предыдущие выводы о значительном снижении качества предсказательной модели при тестировании на материале другого жанра.

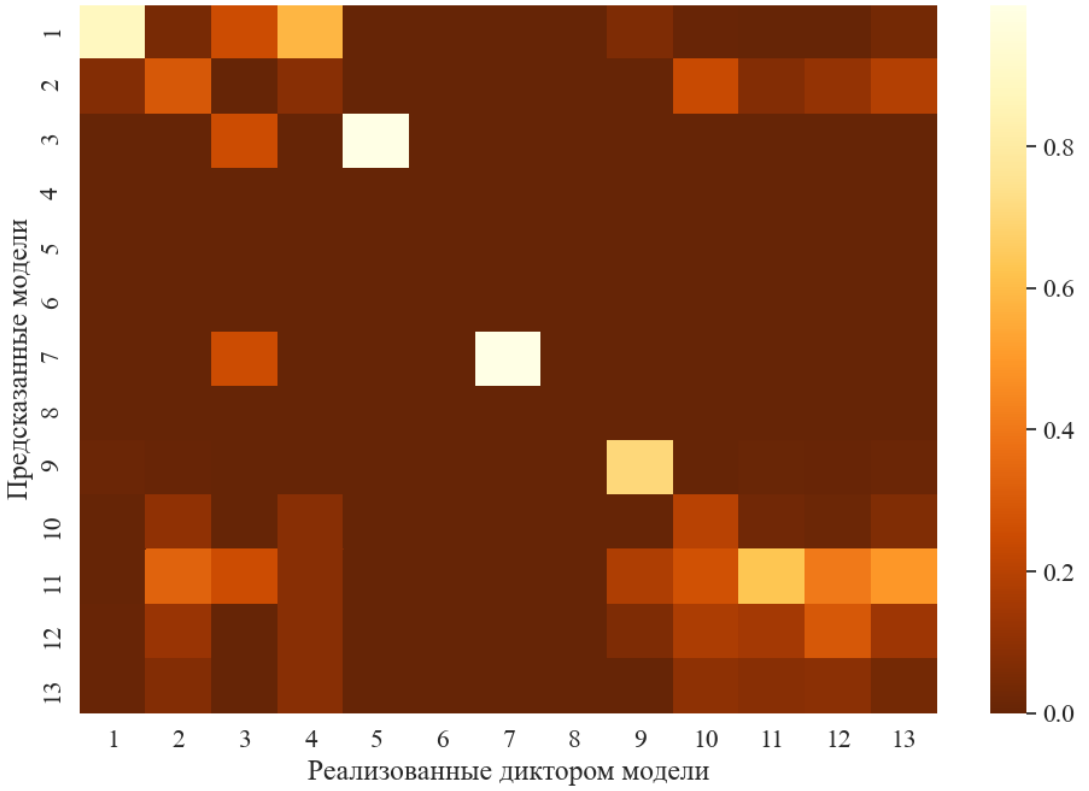


Рисунок 4. Матрица смешения для дикторонезависимой модели, обученной на художественных текстах и протестированной на новостных.

4.7.2 Примеры ошибок

Если детально рассматривать ошибки, допущенные системой при расстановке интонационных центров[4], то можно увидеть, что они в первую очередь возникают там, где дикторы в целях добавления дополнительной

[4] В данном случае рассматривается вывод дикторонезависимой системы, обученной и протестированной на художественных текстах, однако суть ошибок в большинстве своем одинакова для всех моделей.

интонационной выделенности переносят интонационный центр с последнего слова на предшествующие ему (жирным выделен реализованный диктором интонационный центр, курсивом — предсказанный моделью; в скобках указан дополнительный контекст из соседних синтагм):

- “вызывают не **скорую** *помощь* (а пожарную команду)” (В данном предложении перенос осуществили семь дикторов из восьми)
- “и на **работе** его *уважают*” (Перенос осуществили восемь дикторов из восьми)
- “а тут ему **музыка** *помешала*!” (Перенос осуществили восемь дикторов из восьми)
- “(ниже) на целых **две** *головы*!..” (Перенос осуществили два диктора из восьми)
- “**очень** похож на *дедушку*!” (Перенос осуществили пять дикторов из восьми)
- “**хочешь** в *буфет*?” (Перенос осуществили шесть дикторов из восьми)
- “**я** *думал* (что ничего не пойму...)” (Перенос осуществил один диктор из восьми)
- “и **они** *пошутили.*” (Перенос осуществили восемь дикторов из восьми)
- “я **всегда** *говорю* (что нет безвыходных положений)” (Перенос осуществили шесть дикторов из восьми)

Дополнительное интонационное выделение здесь появляется обычно в целях контраста, фокуса или противопоставления, иногда — в целях эмфазы, иногда обусловлено прагматически (“**хочешь** в буфет” вместо “хочешь в **буфет**” — первый вариант прочтения подразумевает скорее уточнение, действительно ли собеседник хочет пойти, в то время как второй — хочет собеседник в буфет или в другое место), иногда происходит из знаний о сочетаемости и частотности языковых единиц (в словосочетании “скорая помощь” центр ставится на слово “скорая”, вероятно, потому, что оно часто употребляется самостоятельно). Формализация данных значений представляет собой отдельную сложную задачу; без нее, однако, нейросетевая модель в любом случае будет часто

допускать такие ошибки. При этом, как можно видеть из примечаний к каждому примеру, зачастую такие ошибки весомы, поскольку они допускаются в местах, где большинство дикторов корпуса осуществили перенос интонационного центра.

Но все же встречаются и более веские ошибки в предсказании интонационного центра, как, например, когда он не переносится моделью со служебного слова, являющимся последним в синтагме:

- “а потом **шея** *его* (и затылок, и уши)”
- “и все о **том** *же*”

Избежать подобных ошибок можно либо с помощью эмбеддингов слов, если для них доступно достаточное количество обучающего материала, либо с помощью обработки клитик: если подобные комплексы, как “о том же”, будут учитываться системой как единое фонетическое слово, предсказанным интонационным центром станет то слово, что несет основное ударение в фонетическом комплексе.

Перенос же интонационного центра моделью там, где диктором это осуществлено не было, выглядят обоснованным вариантом прочтения:

- “и с *Лорой* **тоже.**”
- “глаза ее *обрадованно* **засияли.**”
- “так и *должно* **быть.**”
- “он тебе *очень* **идет.**”
- “они *перебивали* друга **друга**.”

Примеры показывают, что модель, несмотря на отсутствие формального семантического анализа, способна обучаться стратегиям переноса интонационного центра и порождать более эмоциональные варианты прочтения.

Примерами для ошибок в определении интонационных моделей, описанных в п. 4.7.1, могут послужить, например, следующие предложения, где на месте дикторской модели 02 (завершенность с выделенностью) была предсказана интонация незавершенности (первое число в квадратных скобках

означает реализованную диктором модель, второе число — предсказанную; две черты указывают на синтагматическую границу):

- “отец не **пропел** [2, 11], || а как-то почти **прошептал** [2, 11]”
- “она говорила **бодро** [2, 11], (и даже весело)”
- “он упрямо пытался **сводить** [2, 11] || **мирить** [2, 11], || селил вместе на **даче** [2, 11]”

Реализация модели 11 в таких предложениях возможна, поэтому смешение моделей 02 и 11 здесь не представляется серьезной ошибкой; впрочем, вероятно, имеет смысл разделять союзы на несколько категорий и формализовывать данное разделение в виде признака, чтобы противопоставительные союзы, такие, как “а”, учитывались в качестве контекста и повышали вероятность появления модели 02, используемой иногда для противопоставления одной синтагмы другой (как в первом примере).

Предсказание модели 01 (нейтральной завершенности) на месте восклицательной модели 04 происходило как в предложениях с эксплицитным указанием на восклицательность в виде наличия соответствующего знака, так и в предложениях, где эмоциональная коннотация была результатом дикторской интерпретации:

- “пусть **едет** [4, 1].”
- “(говорила про нее Дмитриеву:) **ханжа** [4, 1].”
- “(не волнуйся) и спи **спокойно** [4, 1].”
- “не **сердись** [4, 1].”
- “тебе звонили из **ИМКОИНа** [4, 1]!”
- “вполне взрослый **парень** [4,1]!”

Одним из назначений модели 04 является также оформление просьб — и это можно увидеть из приведенных примеров (примеры 3 и 4). Несмотря на то, что признак, указывающий на императивную форму слова, был использован в модели, его вес, вероятно, оказался слишком мал (ввиду невысокой частотности императивов), и потому не сработал во всех нужных случаях.

Случаи ошибочного определения модели 07 на месте модели 06 были найдены только для единичных вопросительных предложений, поэтому данная ошибка, которая бросается в глаза при анализе матрицы смешения и теоретически может быть перцептивно значимой при синтезе речи, все же не кажется слишком крупной, а является скорее статистическим выбросом:

- “ты **слышал** [6, 7]?”
- “**ну** [6,7]?”

Предсказание нейтральной модели 01 на месте модели 02 (более интонационно выделенного варианта) часто отмечалось в предложениях с оценочными прилагательными или предложениях, описывающих эмоциональное состояние субъекта:

- “а он приходил в **ярость** [2, 1].”
- “был действительно человек **могучий** [2, 1].”
- “Лена была **удивлена** [2, 1].”
- “нет ничего **отвратительнее** [2, 1].”
- “это нечто **прекрасное** [2, 1].”
- “почти **немыслимо** [2, 1].”
- “Лена держалась **великолепно** [2, 1].”

Дополнительные средства автоматизированного семантического анализа или предсказания интонационно выделенных слов кажутся важным инструментом для решения данной проблемы, поскольку эмбеддингов BERT оказалось недостаточно для ее решения.

Предсказательная модель также не всегда справляется с интонацией недоуменного или риторического вопроса, определяя на месте модели 05 — модель 03:

- “(господи,) какие **проблемы** [5, 3]?”
- “какой же я **взрослый** [5, 3]?”
- “где же простая **логика** [5, 3]?”

На месте модели 05, обозначающей восклицание, системой иногда предсказывается модель 04, что является допустимым вариантом интонирования:

- “на столько **лет** [5, 4]!”
- “такие **надежные** [5, 4]!”
- “столько часов без **обеда** [5, 4]!”

4.8 Сравнение нейросетевого подхода с интонационным аннотатором, работающим на правилах

Как было описано в теоретическом разделе 3, существует множество систем, целью которых является предсказание интонационного оформления текста. Прямое сравнение их с результатами представленного в данном исследовании метода, впрочем, было бы некорректно из-за многочисленных различий: в языке, материале, постановке эксперимента, используемых систем интонационного описания. По этой причине здесь будет представлено только сравнение с интонационном аннотатором, разработанным на кафедре фонетики и методики преподавания иностранных языков СПбГУ И. В. Жарковым и коллегами (Жарков и др., 1995): комплексной системой предсказания интонационного оформления (включая такие его аспекты, как ударения, синтагматические границы, интонационные центры и модели, паузы и пр.). Система работает на правилах (поэтому далее в тексте будет называться “системой на правилах”) и опирается на анализ поверхностной структуры. Поскольку данная система использует интонационную систему Н. Б. Вольской для разметки, а также поскольку были доступны результаты ее работы на корпусе CORPRES, представилось возможным сравнить ее эффективность с эффективностью нейросетевого подхода.

Для сравнения была взята дикторонезависимая модель для художественных текстов; ее выдача для тестовых выборок на всех кросс-валидационных итерациях была сопоставлена с разметкой, порожденной с помощью системы на правилах. Сравнивалась только точность предсказания интонационной модели:

сравнение предсказаний расположения интонационных центров было бы некорректным, т.к. этот аспект интонационного оформления крайне сильно зависит от синтагматических границ, не всегда совпадающих для системы на правилах и нейросетевого метода — система на правилах предсказывает синтагматические границы самостоятельно, а для нейросетевого метода они извлекаются из разметки.

Процедура подсчета точности предсказания интонационных моделей также была модифицирована. Причиной этому было то, что одно и то же предложение зачастую может быть прочитано с разной интонацией, и аннотатор и нейросеть могут предсказывать разные одинаково допустимые варианты интонирования для предложения. Чтобы не сравнивать их с прочтением какого-либо одного диктора, было принято решение засчитывать интонационную модель за правильно предсказанную в том случае, если она была реализована на текущем слове хотя бы одним диктором корпуса[5].

Также было реализовано два условия сравнения результатов:

1) В первом случае **все интонационные модели**, предсказанные методами, сравнивались с теми, что были реализованы дикторами корпуса.
2) Во втором случае сравнение проводилось только в том случае, если **показатель наличия синтагматической границы** после слова **совпадал** для нейросетевого метода и системы на правилах; другими словами, только в том случае, если синтагматическая граница, предсказанная аннотатором, совпадала с синтагматической границей, извлеченной из разметки для диктора тестовой выборки и потому учитываемой нейросетью в качестве признака при предсказании.

Второе условие сравнения позволяло снизить вероятность того, что расхождения в предсказаниях обусловлены различием в постановке синтагматических границ, а не прямой ошибкой предсказательного метода. Тем не менее, оно несколько способствует улучшению значения метрики точности

[5] Важно отметить, что нейросетевая модель, обучаемая на том же корпусе, для которого проводилось тестирование, не получала при этом несправедливого преимущества, поскольку предложения тестовой и обучающей выборки не пересекались.

для обоих методов, поскольку в этом случае учитываются только те синтагмы, где аннотатор поставил границы так же, как и диктор тестовой выборки: вероятно, это зачастую происходило в несколько более однозначных и, в целом, более простых для предсказания интонации предложениях.
Результаты сравнения представлены в Таблице 14.

Таблица 14. Сравнение точности предсказания интонационных моделей, получаемой с помощью нейросетевого подхода (дикторонезависимая модель) и интонационного системы на правилах (на материале художественных текстов).

Предсказательный метод	Точность предсказания интонационных моделей, %	
	Условие 1 (все интонационные модели)	Условие 2 (модели в совпадающих синтагмах)
Нейросетевой	**84,5**	**85,1**
Система на правилах	56,9	61,8

Вне зависимости от условия сравнения значения метрики точности для нейросетевого метода выше, чем для системы на правилах: в первом случае точность нейросетевого метода в 1,5 раза больше, во втором — в 1,4 раза. На основе данных результатов можно сделать вывод, что нейросетевой подход к предсказанию интонации может позволить достичь более высокого уровня качества, чем система на правилах. Разумеется, важно при этом помнить, что нейросетевые модели требуют большого количества размеченных данных для обучения (как правило, значительно большего, чем требуется лингвистам для анализа и разработки алгоритма предсказания).

Заключение

В ходе работы был предложена и протестирована система предсказания интонационного оформления для русского текста. Система основана на нейросетевой архитектуре BiLSTM и предсказывает расположение интонационных центров и интонационные модели для синтагм в терминах системы интонационного описания Н. Б. Вольской. В качестве материала для обучения и тестирования был использован корпус CORPRES, включая содержащиеся в нем художественные и новостные тексты, прочитанные восемью профессиональными дикторами.

Классификационные признаки включали следующие группы: пунктуационную, морфологическую, лексическую, синтаксическую и фонетическую. Эксперименты показали, что использование всех этих признаков оправдано, а их комбинация дает наилучшие результаты; однако эмбеддинги BERT, передававшие лексический аспект предложения, дают малый прирост эффективности, вероятно, в связи с недостаточным количеством материала. Для устранения данной проблемы необходимо увеличить объем обучающей выборки, либо снизить размерность эмбеддингов слов. Среди признаков, оказывающих наибольшее влияние на предсказание расположения интонационного центра, были фонетические, пунктуационные и морфологические. Точность предсказания интонационной модели же повышалась в первую очередь за счет синтаксических, фонетических и пунктуационных признаков.

Предсказание расположения интонационных центров для дикторонезависимой модели на художественных текстах осуществлялось с эффективностью в 88,7%, предсказание интонационной модели — с точностью в 61,2% (или 84,5%, если считать правильной интонационную модель, которую в данном предложении реализовал хотя бы один диктор корпуса). Значения метрик для дикторозависимых моделей значительно варьируются; однако, можно сказать, что эффективность предсказания интонационных моделей для

дикторозависимых моделей выше, а точность предсказания расположения центров меньше. Дальнейшим этапом улучшения системы предсказания может стать ее разделение на два компонента, один из которых будет обучаться на дикторонезависимом материале, а другой — на дикторозависимом.

Интонация новостных текстов предсказывается системой с большей эффективностью, чем интонация художественных, и в случае дикторозависимых, и в случае дикторонезависимых моделей. Это связано в первую очередь с ее меньшей вариативностью, большей нейтральностью и большей предсказуемостью; в связи с этим, в условиях разнородности обучающей и тестовой выборок системы, обученные на новостных текстах, показывали более плохие результаты, чем системы, обученные на художественных текстах.

Точность предсказания интонационных моделей для описанной в работе системы была сравнена с точностью системы на правилах, разработанной И. В. Жарковым и коллегами на кафедре фонетики и методики преподавания иностранных языков СПбГУ; нейросетевая система продемонстрировала улучшение по сравнению с системой на правилах в 1,4 раза.

Список литературы

1. Бондарко Л. В. [и др.]. Фонетика спонтанной речи / Л. В. Бондарко [и др.], Издательство Ленинградского университета, 1988.
2. Брызгунова Е. А. Практическая фонетика и интонация русского языка: пособие для преподавателей, занимающихся с иностранцами / Е. А. Брызгунова, Издательство Московского Университета, 1963.
3. Вольская Н. Б., Скрелин П. А. Система интонационных моделей для автоматической интерпретации интонационного оформления высказывания: функциональные и перцептивные характеристики // Третий междисциплинарный семинар" Анализ разговорной русской речи" (АРЗ-2009) – 2009. С. 28–40.
4. Жарков И. В., Слободянюк С. Л., Светозарова Н. Д. Автоматический акцентно-интонационный транскриптор произвольного русского текста / Бюллетень фонетического фонда русского языка. 1995. № 3. С. 58–70.
5. Ковтунова И. И. Современный русский язык: Порядок слов и актуальное членение предложения / И. И. Ковтунова, Просвещение, 1976.
6. Кодзасов С. В., Кривнова О. Ф. Общая фонетика / С. В. Кодзасов, О. Ф. Кривнова, Издательство Российского Гуманитарного Государственного Университета, 2001.
7. Кривнова О. Ф., Князев С. В., Моисеева Е. В. Исследования просодического членения звучащего текста на материале русского языка // Вестник Московского университета. Серия 9. Филология. – 2016. – №. 4.
8. Лобанов Б. М. Алгоритм сегментации текста на синтаксические синтагмы для синтеза речи // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Междунар. конф. «Диалог». М, 2008.
9. Николаева Т. М. Семантика акцентного выделения / Т. М. Николаева, М.: "Наука", 1982.

10. Пронникова Н. В. К вопросу о функциях интонации / Фундаментальные исследования. 2014. № 9 (5).
11. Рыбин С. В. Синтез речи / С. В. Рыбин, Санкт-Петербург: Университет ИТМО, 2014.
12. Светозарова Н. Д. Интонационная система русского языка 1982.
13. Светозарова Н. Д., Штерн А. С. Ключевые и фонетически выделенные слова текста // Экспериментальная фонетика. Автоматическое распознавание и синтез речи: сборник статей. – 1989. – С. 157.
14. Слюсарь Н. А. На стыке теорий / Н. А. Слюсарь, Общество с ограниченной ответственностью "Книжный дом ЛИБРОКОМ", 2009.
15. Черемисина-Ениколопова Н. В. Русская интонация: поэзия, проза, разговорная речь / Н. В. Черемисина-Ениколопова, Русский язык, 1989.
16. Янко Т. Е. Интонационные стратегии русской речи в сопоставительном аспекте / Т. Е. Янко, Litres, 2017.
17. Abadi M. [и др.]. TensorFlow: Large-scale machine learning on heterogeneous systems / M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, [и др.]., 2015.
18. Annamoradnejad I., Zoghi G. Colbert: Using bert sentence embedding for humor detection // arXiv preprint arXiv:2004.12765. 2020.
19. Bach N. X., Duy T. K., Phuong T. M. A POS tagging model for vietnamese social media text using BiLSTM-CRF with rich features // Pacific Rim International Conference on Artificial Intelligence. – Springer, Cham, 2019. – С. 206-219.
20. Brenier J. M., Cer D. M., Jurafsky D. The detection of emphatic words using acoustic and lexical features // Ninth European Conference on Speech Communication and Technology. – 2005.
21. Bulyko I., Ostendorf M. Joint prosody prediction and unit selection for concatenative speech synthesis IEEE // 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221). IEEE, 2001. – Т. 2. – С. 781-784.

22. Burtsev M. [и др.]. Deeppavlov: Open-source library for dialogue systems // Proceedings of ACL 2018, System Demonstrations. – 2018. – С. 122-127.
23. Chen K., Hasegawa-Johnson M., Cohen A. An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model IEEE // 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. – IEEE, 2004. – Т. 1. – С. I-509.
24. Chen Q., Zhuo Z., Wang W. Bert for joint intent classification and slot filling // arXiv preprint arXiv:1902.10909. 2019.
25. Chollet F. Keras: The python deep learning library // Astrophysics Source Code Library. 2018.
26. Devlin J. [и др.]. Bert: Pre-training of deep bidirectional transformers for language understanding // arXiv preprint arXiv:1810.04805. 2018.
27. Gravano A., Hirschberg J. Turn-taking cues in task-oriented dialogue // Computer Speech & Language. 2011. № 3 (25). С. 601–634.
28. Guo H. [и др.]. Exploiting syntactic features in a parsed tree to improve end-to-end TTS // arXiv preprint arXiv:1904.04764. 2019.
29. Hayashi T. [и др.]. Pre-Trained Text Embeddings for Enhanced Text-to-Speech Synthesis. // INTERSPEECH. – 2019. – С. 4430-4434.
30. Helander E. E., Nurminen J. A novel method for prosody prediction in voice conversion IEEE //2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. – IEEE, 2007. – Т. 4. – С. IV-509-IV-512.
31. Hirschberg J. Using text analysis to predict intonational boundaries //Second European Conference on Speech Communication and Technology. – 1991.
32. Hirschberg J. Pitch accent in context predicting intonational prominence from text // Artificial Intelligence. 1993. № 1–2 (63). С. 305–340.
33. Hirschberg J., Prieto P. Training intonational phrasing rules automatically for English and Spanish text-to-speech // Speech Communication. 1996. № 3 (18). С. 283–292.
34. Hovy D. [и др.]. Analysis and modeling of "focus" in context // Interspeech. –2013. – С. 402-406.

35. Ingulfsen T. Influence of syntax on prosodic boundary prediction / T. Ingulfsen, University of Cambridge, Computer Laboratory, 2004.
36. Ishihara S. On the (lack of) correspondence between syntactic clauses and intonational phrases [Электронный ресурс]. URL: https://ling.auf.net/lingbuzz/005414
37. Kandel I., Castelli M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset // ICT express. 2020. № 4 (6). С. 312–315.
38. Klimkov V. [и др.]. Phrase Break Prediction for Long-Form Reading TTS: Exploiting Text Structure Information. // Interspeech. – 2017. – С. 1064-1068.
39. Kocharov D., Kachkovskaia T., Skrelin P. Prosodic boundary detection using syntactic and acoustic information // Computer Speech & Language. – 2019. – Т. 53. – С. 231-241.
40. Koehn P. [и др.]. Improving intonational phrasing with syntactic information // 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100). – IEEE, 2000. – Т. 3. – С. 1289-1290.
41. Kuratov Y., Arkhipov M. Adaptation of deep bidirectional multilingual transformers for russian language // arXiv preprint arXiv:1905.07213. 2019.
42. Li J., Nenkova A. Fast and accurate prediction of sentence specificity // Proceedings of the AAAI Conference on Artificial Intelligence. – 2015. – Т. 29. – №. 1.
43. Liu R. [и др.]. Improving Mongolian Phrase Break Prediction by Using Syllable and Morphological Embeddings with BiLSTM Model // Interspeech. – 2018. – С. 57-61.
44. Louw J. A., Moodley A. Speaker specific phrase break modeling with conditional random fields for text-to-speech // 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech). – IEEE, 2016. – С. 1-6.

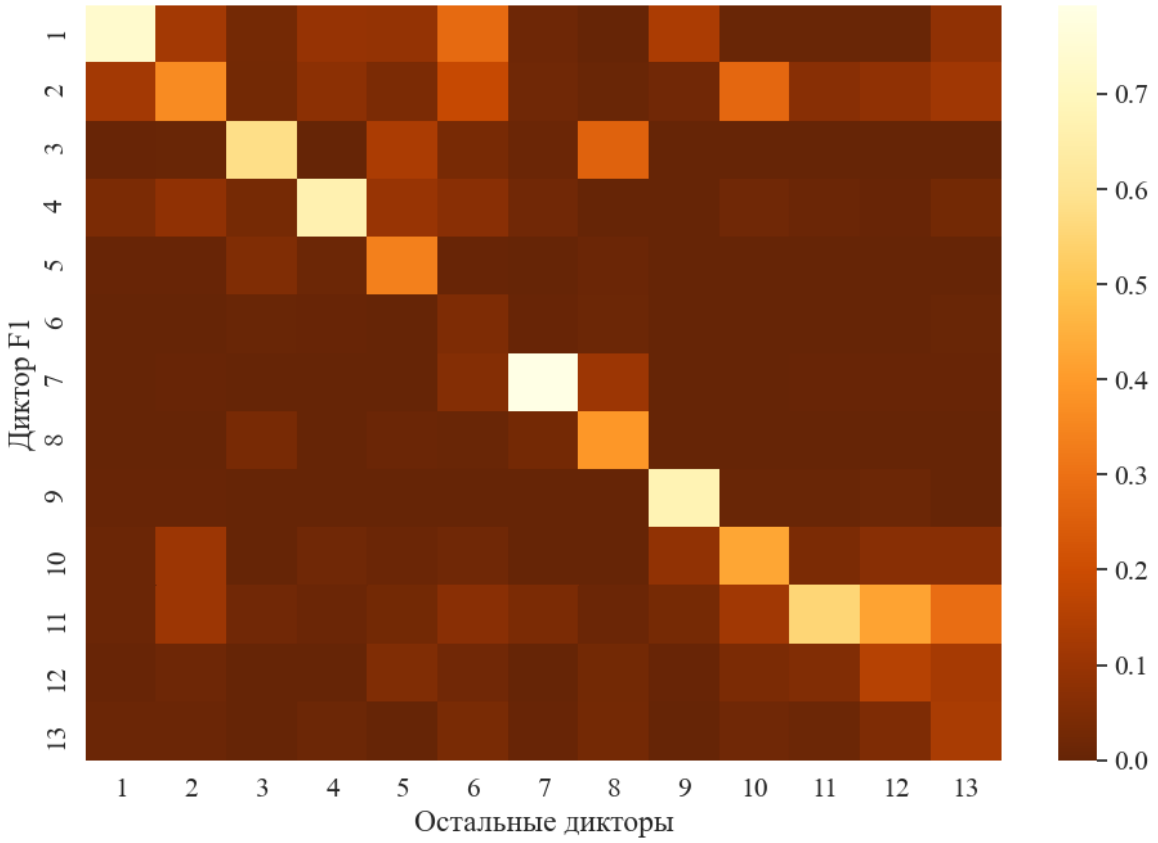
45. Menshikova A., Kocharov D. Prosodic Boundaries Prediction in Russian Using Morphological and Syntactic Features // Conference on Artificial Intelligence and Natural Language. – Springer, Cham, 2019. – С. 126-135.
46. Mishra T., Kim Y., Bangalore S. Intonational phrase break prediction for text-to-speech synthesis using dependency relations // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2015. – С. 4919-4923.
47. Nakajima H., Mizuno H., Sakauchi S. Emphasized accent phrase prediction from text for advertisement text-to-speech synthesis // Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing. – 2014. – С. 170-177.
48. Nespor M., Vogel I. Prosodic phonology / M. Nespor, I. Vogel, De Gruyter Mouton, 2012.
49. Nivre J. [и др.]. Universal Dependencies 2.1 2017.
50. Obin N., Lanchantin P. Symbolic modeling of prosody: From linguistics to statistics // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2015. № 3 (23). С. 588–599.
51. Ostendorf M., Price P. J., Shattuck-Hufnagel S. Combining statistical and linguistic methods for modeling prosody // ESCA Workshop on Prosody. – 1993.
52. Ostendorf M., Price P. J., Shattuck-Hufnagel S. The Boston University radio news corpus // Linguistic Data Consortium. 1995. С. 1–19.
53. Pascual S., Bonafonte A. Prosodic break prediction with RNNs // International Conference on Advances in Speech and Language Technologies for Iberian Languages. – Springer, Cham, 2016. – С. 64-72.
54. Pennebaker J. W. [и др.]. The development and psychometric properties of LIWC2015. 2015.
55. Read I., Cox S. Stochastic and syntactic techniques for predicting phrase breaks // Computer Speech & Language. 2007. № 3 (21). С. 519–542.

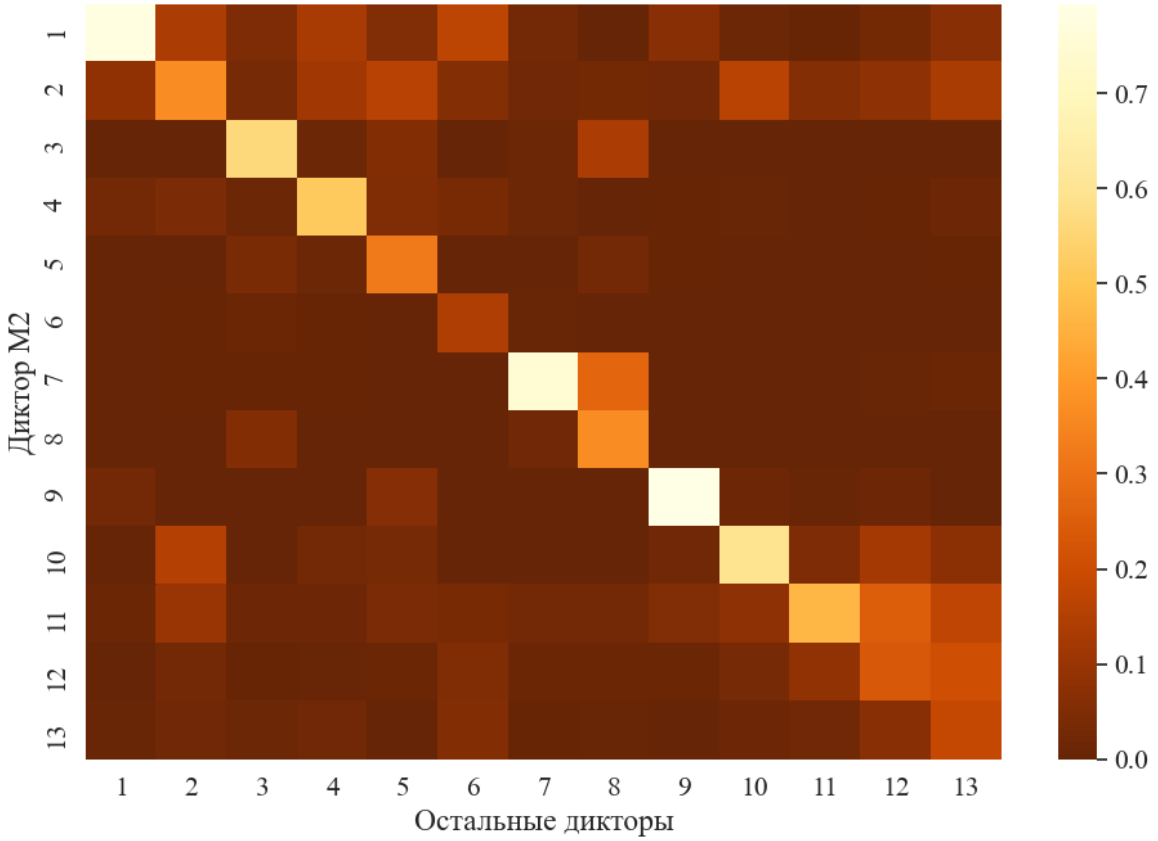
56. Ren Y. [и др.]. Fastspeech 2: Fast and high-quality end-to-end text to speech // arXiv preprint arXiv:2006.04558. 2020.
57. Rendel A. [и др.]. Using continuous lexical embeddings to improve symbolic-prosody prediction in a text-to-speech front-end //2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2016. – С. 5655-5659.
58. Rendel A. [и др.]. Weakly-Supervised Phrase Assignment from Text in a Speech-Synthesis System Using Noisy Labels // Interspeech. – 2017. – С. 759-763.
59. Rosenberg A., Fernandez R., Ramabhadran B. Phrase boundary assignment from text in multiple domains // Thirteenth Annual Conference of the International Speech Communication Association. – 2012.
60. Schmid H., Atterer M. New statistical methods for phrase break prediction // COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics. – 2004. – С. 659-665.
61. Selkirk E. 14 The Syntax-Phonology Interface // The handbook of phonological theory. 2011. С. 435.
62. Shen J. [и др.]. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2018. – С. 4779-4783.
63. Silverman K. [и др.]. ToBI: A standard scheme for labeling prosody // Proceedings of the Second International Conference on Spoken Language Processing. – 1992. – С. 867-879.
64. Skrelin P. [и др.]. Corpres // International Conference on Text, Speech and Dialogue. – Springer, Berlin, Heidelberg, 2010. – С. 392-399.
65. Sloan R. [и др.]. Prosody prediction from syntactic, lexical, and word embedding features //10th ISCA Speech Synthesis Workshop. – 2019.
66. Steedman M. J. Intonation and syntax in spoken language systems // Technical Reports (CIS). 1989. С. 838.

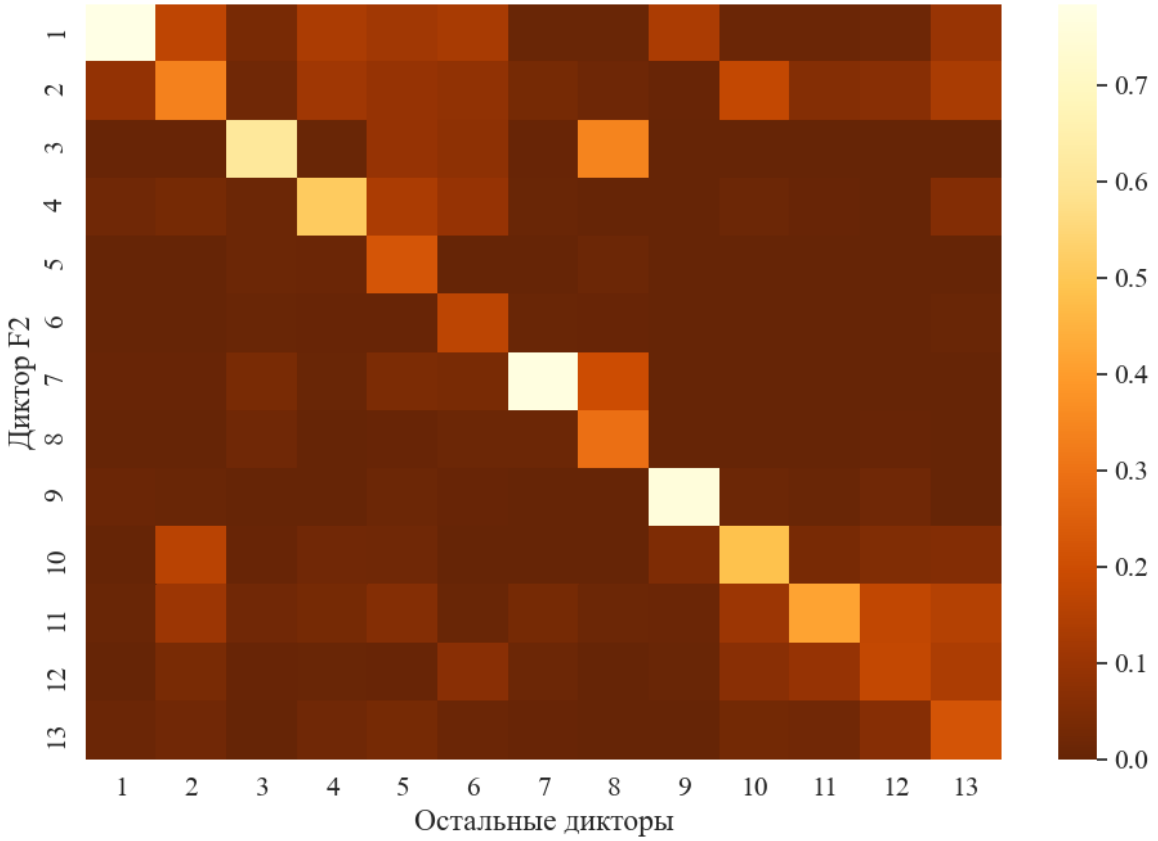
67. Sun X., Applebaum T. H. Intonational phrase break prediction using decision tree and n-gram model // Seventh European Conference on Speech Communication and Technology. – 2001.
68. Tachibana H., Uenoyama K., Aihara S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention IEEE // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2018. – С. 4784-4788.
69. Tamburini F., Caini C. An automatic system for detecting prosodic prominence in American English continuous speech // International Journal of speech technology. 2005. № 1 (8). С. 33–44.
70. Taylor P. Text-to-speech synthesis / P. Taylor, Cambridge university press, 2009.
71. Tepperman J., Nava E. Where should pitch accents and phrase breaks go? A syntax tree transducer solution // Twelfth Annual Conference of the International Speech Communication Association. – 2011.
72. Tyagi S. [и др.]. Dynamic prosody generation for speech synthesis using linguistics-driven acoustic embedding selection // arXiv preprint arXiv:1912.00955. 2019.
73. Vaswani A. [и др.]. Attention is all you need // arXiv preprint arXiv:1706.03762. 2017.
74. Volskaya N., Kachkovskaia T. Prosodic annotation in the new corpus of Russian spontaneous speech CoRuSS // Proceedings of Speech Prosody. – 2016. – Т. 2016.
75. Wang Y. [и др.]. Tacotron: Towards end-to-end speech synthesis // arXiv preprint arXiv:1703.10135. 2017.
76. Xu G. [и др.]. Improving Prosody Modelling with Cross-Utterance Bert Embeddings for End-to-End Speech Synthesis // ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2021. – С. 6079-6083.

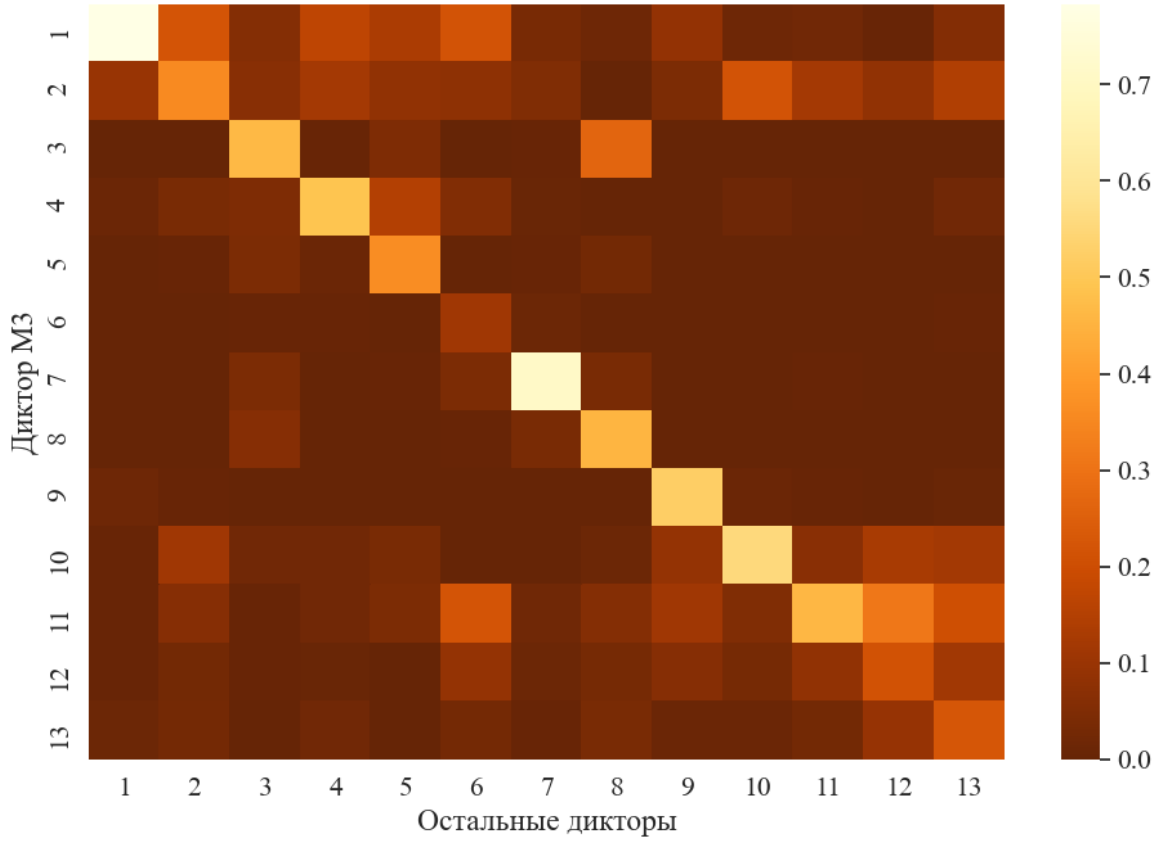
Приложения

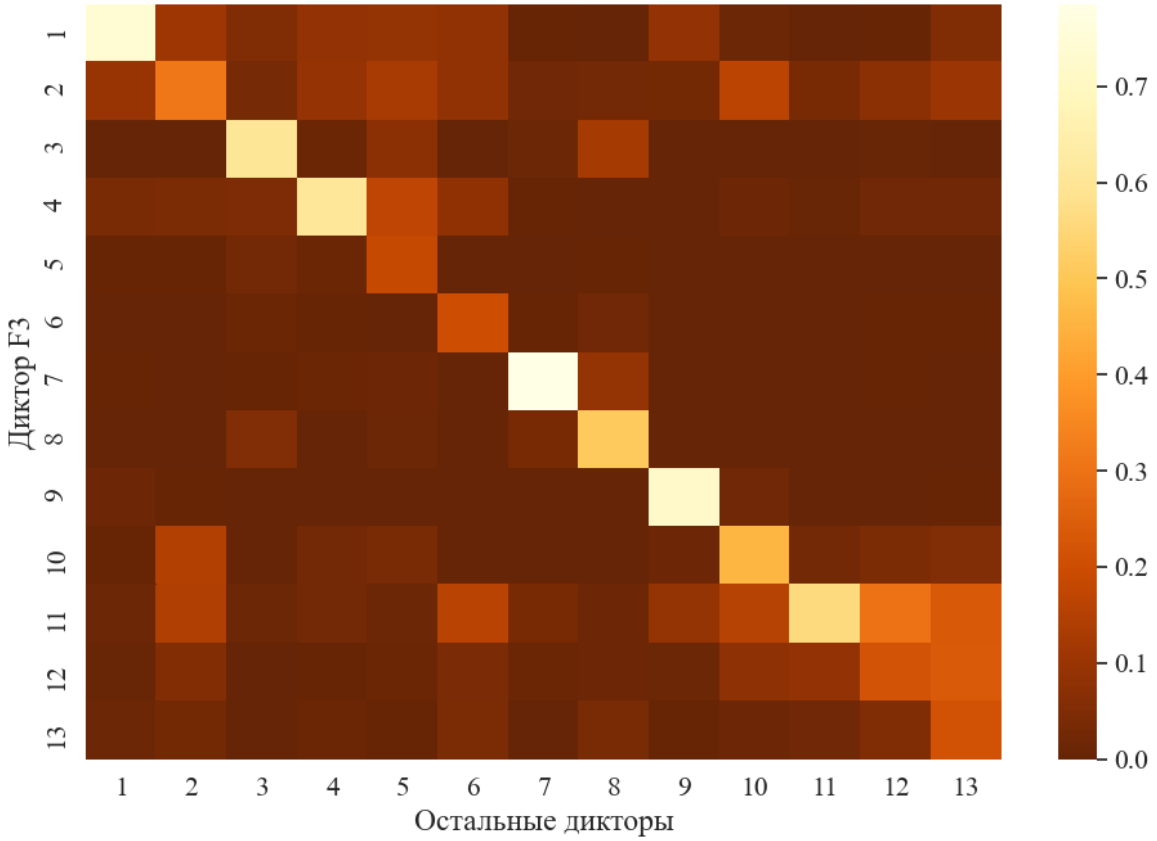
Приложение 1: матрицы смешения для всех дикторов корпуса

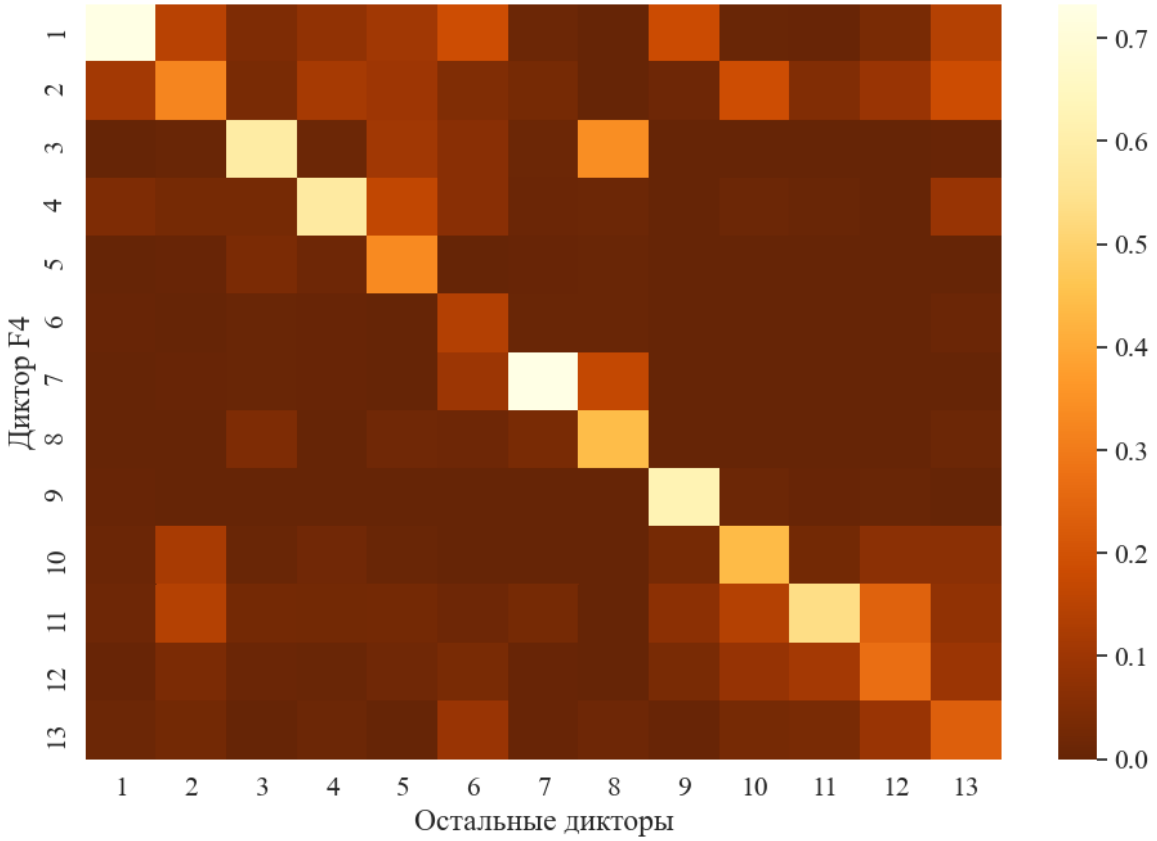


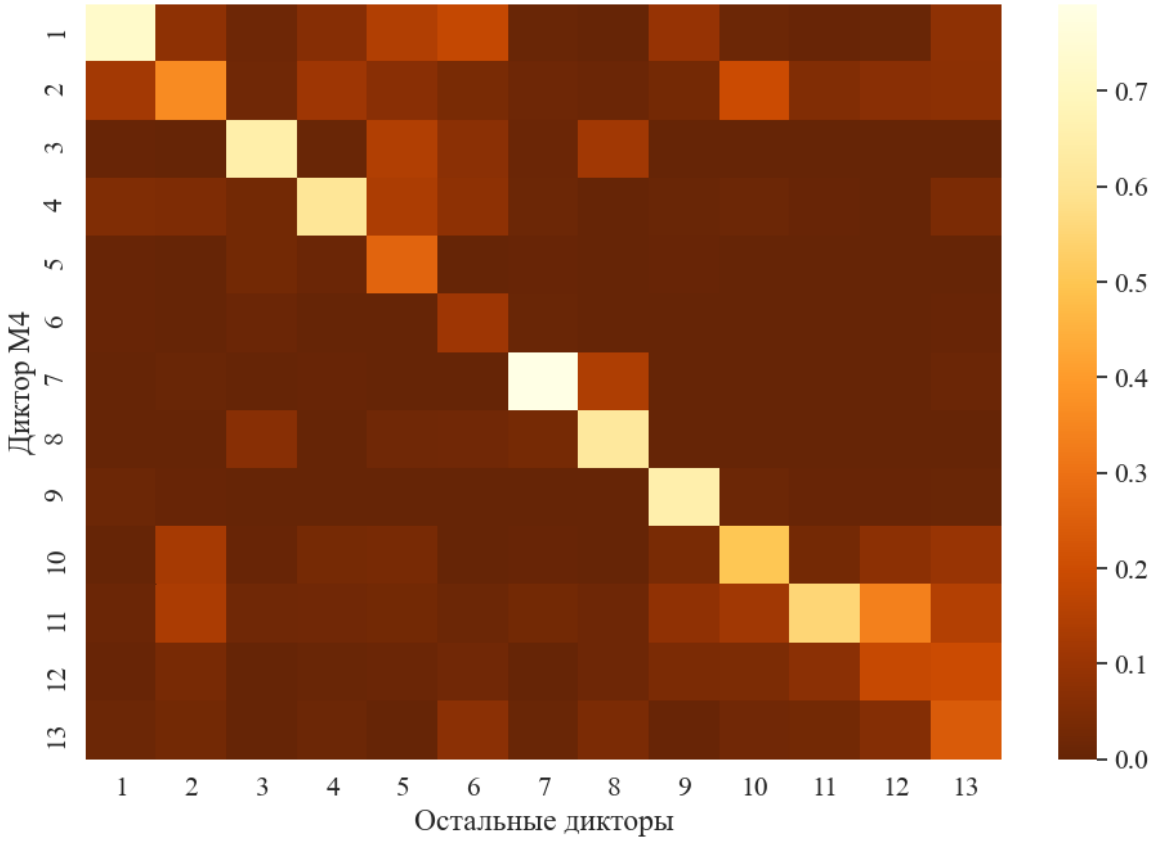

Диктор M2
1
2
3
4
5
6
7
8
9
10
11
12
13
Остальные дикторы
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0


Диктор F2
1
2
3
4
5
6
7
8
9
10
11
12
13
Остальные дикторы
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0


Диктор М3
1
2
3
4
5
6
7
8
9
10
11
12
13
Остальные дикторы
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0


Диктор F3
1
2
3
4
5
6
7
8
9
10
11
12
13
Остальные дикторы
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0


Диктор F4
1
2
3
4
5
6
7
8
9
10
11
12
13
Остальные дикторы
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0


Диктор M4
1
2
3
4
5
6
7
8
9
10
11
12
13
Остальные дикторы
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0