С.А. Шаров, С.О. Савчук

ТИПОЛОГИЯ ТЕКСТОВ ДЛЯ ПРЕДСТАВИТЕЛЬНОГО КОРПУСА

> Животные подразделяются на: а) принадлежащих Императору, б) бальзамированных, в) прирученных, г) молочных поросят, д) сирен, е) сказочных, ж) бродячих собак, з) включенных в настоящую классификацию, и) буйствующих, как в безумии, к) неисчислимых, л) нарисованных очень тонкой кисточкой из верблюжьей шерсти, м) прочих, н) только что разбивших кувшин, о) издалека кажущихся мухами.
>
> *Х.Л. Борхес.*
> *Аналитический язык Джона Уилкинса.*
> *Пер. Е. Лысенко.*

1. Введение

Очевидно, что при создании представительного корпуса требуется обеспечить максимально широкое покрытие различных типов текстов и функциональных стилей. Однако для этого надо иметь список таких типов текстов в той области, к которой относится корпус. Если нашей задачей является создание Национального Корпуса, т.е. представительного корпуса для всего русского языка[1], то нам необходимо составить список *всех* типов текстов, которые существуют в русской культуре. Если мы отталкиваемся от традиционных наименований типов текстов и строим линейный список всех типов, то мы получаем классификацию, подобную Борхесовской, ср. например, список газетных жанров: Analysis, Biography, Commentary, Criticism, Economics, Editorial,

[1] *Шаров С.А.* Представительный корпус русского языка в контексте мирового опыта // Научно-техническая информация. Сер. 2. 2003. №6. С. 8–18.

Finance[1]... Очевидно, что следует структурировать такие списки, чтобы было можно описывать тексты корпуса по разным параметрам.

В современных лингвистических исследованиях ведется поиск адекватного основания для классификации текстов, в качестве которого рассматриваются типы коммуникативных ситуаций, факторы социальной дифференциации языка (социолингвистика), факторы членения литературного языка[2], стилеобразующие[3] и жанрообразующие факторы[4]. Однако несмотря на существование обширной литературы по данному вопросу, единой типологии текстов пока не разработано[5]. Предлагаемые описания либо слишком обобщенны, либо содержат чрезмерно подробную классификацию отдельных разновидностей текстов, например текстов художественной прозы, журналистики, описательных текстов[6].

Англоязычная литература также содержит множество предложений по классификации текстов; некоторые из них прямо отно-

---

[1] *Santini M.* Text Typology and Statistics. Explorations in Italian Press Subgenres // Italian Journal of Linguistics / Rivista di linguistica. Vol. 13. Issue 2. 2001. P. 339–374. URL: **http://www.itri.brighton.ac.uk/~Marina.Santini/articolo_bertinetto.pdf**.

[2] *Лаптева О.А.* Теория современного русского литературного языка. М., 2003.

[3] *Кожина М.Н.* Стилистика русского языка. М., 1993.

[4] *Шмелева Т.В.* Повседневная речь как лингвистический объект // Русистика сегодня: Функционирование языка: лексика и грамматика. М., 1993.

[5] *Дементьев В.В.* Изучение речевых жанров: обзор работ в современной русистике // Вопросы языкознания. 1997. № 1. С. 109–121.

[6] *Чебанов С.В., Мартыненко Г.Я.* Семиотика описательных текстов: Типологический аспект. СПб., 1999.

сятся к составлению корпусов[1], который ссылается на эту литературу как на «джунгли». Принимая во внимание англоязычные рекомендации, необходимо иметь в виду, что среда использования языка оказывает существенное влияние на классификацию текстов. Так, очевидное несоответствие между российской и англоязычной культурами выражено и на уровне жанров. Такие жанры, как выделяемые в Брауновском Корпусе Foundation Reports или Popular Lore с трудом допускают даже адекватный перевод на русский[2]. Возможные переводные эквиваленты, такие как Отчеты фондов или Самодеятельная литература, не являются терминами в русской стилистической традиции и не соответствуют устойчивым и частотным типам текстов на русском языке.

В данной работе мы предлагаем базовый набор параметров для осмысленной классификации *произвольных* текстов, функционирующих в современной русской культуре, в терминах закрытой классификации, не использующей категории М по Борхесу. Затем мы приводим примеры описания отдельных типов текстов в терминах предложенных классов.

2. Система базовых признаков

Предлагаемая классификация использует две наиболее известные модели для описания текстов, хранимых в корпусах. Первая модель была разработана в рамках Text Encoding Initiative

---

[1] *Lee D.* Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle // Language Learning & Technology. Vol. 5, № 3. September 2001. P. 37–72 / URL: **http://llt.msu.edu/vol5num3/pdf/lee.pdf**

[2] *Kučera H., Francis W.N.* Computational Analysis of Present-day American English. Providence, 1967.

(TEI) и описывает профиль текста[1]. Вторая модель была разработана в рамках системы рекомендаций European Advisory Group on Language Engineering Standards (EAGLES), одна из рекомендаций посвящена типологии текстов[2]. Поскольку обе модели основаны на логических свойствах коммуникации, они могут быть адаптированы для описания русского дискурса[3], где сравниваются возможности TEI и EAGLES и описывается то, как можно выразить категории EAGLES через теги TEI).

Дж. Синклер выделяет два класса параметров описания текстов: внешние (**E**), внеязыковые параметры, которые могут повлиять на структуру или содержание текста, и внутренние (**I**) параметры, отражающие свойства языка, используемого в тексте. Он выделяет три группы **E**-параметров:

1) **E1 (origin)** – параметры, относящиеся к созданию текста автором;

2) **E2 (state)** – параметры, относящиеся к внешним признакам текста;

3) **E3 (aims)** – параметры, относящиеся к целям создания текста и его влиянию на аудиторию;

и два основных **I**-параметра:

1) **I1 (topic)** – предметная область текста;

2) **I2 (style)** – стилистические особенности (частично зависящие от E-параметров).

---

[1] *Sperberg-McQueen C.M., Burnard L.* (eds.). Guidelines for Electronic Text Encoding and Interchange, 2001 // URL:
http://www.hcu.ox.ac.uk/TEI/P4X/index.html

[2] *Sinclair J.* Preliminary recommendations on text typology. EAGLES Document EAG-TCWG-TTYP/P, 1996 // URL :
http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html

[3] *Sharoff S.* Towards basic categories for describing properties of texts in a corpus // Proceedings of Language Resources and Evaluation Conference (LREC04). May 2004. Lisbon, Portugal. URL:
http://www.comp.leeds.ac.uk/ssharoff/texts/lrec-04.pdf

К группе **E1** (параметры создания текста) относятся, в первую очередь, время создания текста и информация об авторстве текста, который может быть создан одним человеком, несколькими поименованными соавторами, либо представлять обобщенного автора, например, в случае передовой статьи в газете или руководства пользователя, где личное авторство принципиально не важно. Для личного автора мы должны указать приблизительные известные характеристики: возраст автора на момент написания текста, пол автора и регион происхождения автора. Для региона важна грубая классификация на столичный (Москва и Санкт-Петербург), европейский, сибирский и южный, для возраста – на детский, молодежный, взрослый и пожилой. При этом предполагается, что большинство текстов в нашем корпусе созданы взрослыми авторами без выраженной региональной специфики.

Для описания текста по его внешним признакам (**E2**) предлагается иерархия, отличающаяся от традиционной, в первую очередь, наличием четырех режимов речи: устной, письменной, письменной, предназначенной для произнесения вслух, и электронной коммуникации. Последняя похожа на устную речь спонтанностью порождения (подобно телефонному звонку или очной дискуссии), но она все равно остается письменной, в частности, в ней отсутствует просодическая информация. К текстам для электронной коммуникации мы относим разного рода группы обсуждений (message boards) и чаты (chats). Письменная речь, предназначенная для произнесения вслух (доклады, официальные выступления, драматургия), также является пограничным случаем, но уже по другой причине: это устная речь с полным набором просодической информации, обычно воспринимаемая слушателем в устном режиме, но изначальный режим ее создания – письменный.

Среди параметров собственно письменной речи выделяются печатные издания, подразделяемые на книги, периодику и брошюры, а также переписка разного рода и то, что Дж. Синклер

называет typed – машинописные тексты, например, отчеты, которые отличаются от печатной продукции по формальному признаку (печатные издания выпускаются, как правило, большим тиражом специализированными структурами, издательствами). При этом современные «машинописные» тексты, подготовленные в текстовых редакторах, могут не отличаться по внешнему виду от книг или брошюр. Устную речь можно подразделять на записанную в естественных условиях, в студии и на телефонные разговоры.

Группа параметров (**E3** по классификации Дж. Синклера) разделена на две группы: цели создания текста (**E3.1**) и его влияние на аудиторию (**E3.2**). К параметрам аудитории, которые оказывают существенное влияние на текст, отнесены ее размер, близость аудитории говорящему и ограничения на пол, возраст и уровень образования аудитории. По размеру аудитории речь делится на публичную (более 50 читателей/слушателей, с подклассами сотни, десятки тысяч и миллионы) и частную, в свою очередь подразделяемую на личную (2 участника), небольшую группу (до 5), группу средних размеров (до 20) и коллектив (до 50). По параметру возраста аудитории разграничиваются тексты, предназначенные для взрослых, для детей, для подростков и рассчитанные на смешанную аудиторию. По параметру близости в большинстве случаев публичная аудитория деперсонализирована. Если говорящий/пишущий может описать каждого участника коммуникации, их близость классифицируется по шкале: хорошее личное знакомство, личное знакомство и его отсутствие.

Образование аудитории может измеряться по двум параметрам: 1) знание о конкретном предмете (общее и специальное) и 2) уровень образования (высокий и низкий). Эти параметры взаимно дополняют друг друга: тексты могут быть предназначены для аудитории без специальных знаний о предмете, но предполагают общий высокий уровень образования. Это означает, что такие тексты используют минимум специальной терминологии, но

могут апеллировать к абстрактным категориям. Напротив, другие тексты могут быть предназначены специалистам с невысоким уровнем образования и использовать большое количество специальной терминологии, но не абстрактных рассуждений по данной теме.

Под целями создания текста (**E3.1**) понимается коммуникативная функция текста, в соответствии с которой тексты делятся на предназначенные для:

обсуждения (discussion), подтипы: аргументация, полемика, изложение позиции**;**

рекомендации (recommendation), подтипы: отчеты, предложения, законы, реклама;

развлечения (recreation), сюда входят различные жанры художественной литературы, а также биографические и автобиографические тексты, дневники и мемуары;

образования (instruction), в эту категорию входят учебники, учебные пособия, практические руководства;

информирования (information), только те тексты, целью которых является исключительно предоставление информации и которые не могут быть включены в другие категории.

Верхний уровень списка коммуникативных функций в основном следует рекомендациям Дж. Синклера (с малыми модификациями), но внутренняя структура подкатегорий подверглась существенной переработке на основании опыта кодирования русских текстов, которые отбирались в корпус. Часто встречается вариация с отнесением текстов отдельных типов к той или иной категории. Например, кулинарные рецепты обычно относятся к обучению (instruction, подтип: практические руководства), но некоторые их виды можно классифицировать как рекомендации. Аналогично прогностические тексты: если прогноз погоды создается с целью информирования (предоставление данных), то астрологический прогноз относится к рекомендациям (это можно отметить и по их лингвистическим особенностям).

При построении корпуса не слишком важна глубина кодирования предметной области, затрагиваемой текстом (параметр **I1**), поскольку корпус не является универсальной энциклопедией. Кроме того, общие классификации, подобные УДК, редко применимы к тексту, так как сколько-нибудь значимый отрезок текста может затрагивать несколько предметных областей одновременно. При построении корпуса можно иметь грубую классификацию, выделяющую естественные и гуманитарные науки, прикладные области (например, строительство или транспорт), политику и экономику, искусство и досуг, а также общую тематику (life), свойственную текстам художественной литературы и мемуаристики.

Система кодирования стилистических особенностей текстов (параметры **I2**) пока не разработана. Наиболее очевидным кажется выделение таких отклонений от нейтрального стиля, как формальный (для деловой речи), академический (для научной речи) и просторечный (для разговорной речи). Для художественной литературы может быть полезным деление текстов на представляющие стандартный литературный язык (например, тексты Ю. Трифонова), сниженный язык (тексты Ю. Алешковского, Э. Лимонова), язык с имитацией региональных особенностей («деревенская проза»), выраженно индивидуальный авторский язык, отличный от нормы (Саша Соколов) и др.

3. Классификация в действии

3.1. Примеры кодирования

Любой текст, предназначенный для включения в корпус, должен быть описан в рамках предложенных базовых параметров. В качестве примера текста в области внутренней политики возьмем текст Конституции РФ: это текст, написанный в формальном стиле (I2) в 1993 г. в Москве, авторство обобщенное (corporate) (E1), это письменный текст, опубликованный в виде

книги объемом 9500 слов (E2), предназначенный для очень большой аудитории без ограничений на образование (даже если он не был прочитан подавляющим большинством населения), цель создания: рекомендация, подтип: юридический документ, обладающий правовой функцией.

Некоторые комбинации параметров являются крайне маловероятными: например, книги, написанные в формальном стиле в области естественных наук для массовой женской аудитории с целью развлечения (комбинация формального стиля изложения и развлекательных целей, а также конкретного пола предполагаемой аудитории и тематики естественных наук кажутся маловероятными). Некоторые параметры взаимно исключают друг друга, например, машинописный отчет, предназначенный для миллионной аудитории или личное обсуждение по телевидению. Все же любое *разумное* сочетание параметров должно быть при возможности представлено в корпусе несколькими текстами.

3.2. Сравнение с традиционными списками типов текстов

«Многомерная» характеристика текстов по предложенным параметрам является более удобным инструментом для оценки представительности корпуса, чем списки традиционных типов текстов, которые, как правило, выделяются сразу на нескольких основаниях (предметная область, цель создания, стиль текста и т.д.).

Так, если мы определяем учебник как тип текста и включаем 50 учебников в наш корпус, мы все равно не можем ничего сказать о представительности корпуса, поскольку существует великое множество учебников, отличающихся в терминах предложенной классификации по предметной области (учебники по физике, философии, иностранному языку или стройматериалам, это параметр I1), стилю (I2, некоторые учебники пишутся в нейтральном стиле, другие в академическом), возрасту, образованию и уровню подготовки аудитории (E3.1, учебники для школьников или студентов, для новичков и продолжающих).

Учитывая, что корпус не является энциклопедией, не имеет смысла включать в корпус учебники по каждому академическому предмету (I1), но зато полезно включить тексты одной предметной области, ориентированные на разные аудитории и имеющие разные цели (учебник/научная (академическая) статья/научно-популярная статья или книга) (E3).

В то же время рассматриваемая система признаков не противоречит принятому в стилистике и типологии текстов членению на функциональные сферы и типы текстов (жанры), а соотносится с ним. Приведем описание некоторых интуитивно очевидных типов текстов в терминах предложенной классификации.

Административно-канцелярские тексты для внутреннего пользования: протокол, акт, отчет о командировке, докладная записка, служебная записка, справка, заявление и др.

Такие тексты имеют, как правило, формальный стиль (степень формальности может варьироваться), существуют в машинописной форме, предназначены для частной аудитории, в основном невелики по объему, имеют личное или корпоративное авторство. Такие тексты создаются с целью информирования. Отдельное подмножество составляют документы судопроизводства (протокол судебного заседания, обвинительное заключение, частное определение, судебное решение), которые имеют юридическую тематику.

Деловые документы для внешнего пользования, или так называемая деловая и коммерческая корреспонденция.

Эта группа отличается от предыдущей тем, что относится к другому виду письменных текстов деловой переписке. Она не столь однородна с точки зрения целей: наряду с информированием, целью создания текстов может быть и рекомендация (письмо-приглашение, письмо-предложение), и обсуждение (в случае решения спорных вопросов: ответ на предложение, рекламация, ответ на рекламацию и пр.).

Собственно научные, или академические тексты: монография, диссертация, статья, тезисы, доклад, рецензия, ответ на рецензию.

Такие тексты отличаются стилем (академический), у них есть явный автор(ы), тексты предназначены для образованных профессионалов, коммуникативная функция – обсуждение. Варьируется размер текста, вид публикации (книги или typed) и размер аудитории.

Научно-справочные тексты: словарь, энциклопедия, справочник, каталог.

У таких текстов обычно обобщенный автор, функция – информирование, подтип справка, аудитория – заинтересованная, часто очень большая. От профессиональной подготовки и возраста аудитории зависит тип справочного издания, например, детская энциклопедия (которая может иметь целью и развлечение), справочник школьника по биологии и справочник терапевта.

Научно-популярные тексты: очерк, книга (монография), статья, лекция.

Отличие от монографии или собственно научной статьи – в аудитории, она большая и заинтересованная (но не профессиональная) и цели: в научно-популярных текстах не только обсуждаются научные проблемы, но и сообщается занимательная информация (т.е. они создаются в том числе с целью развлечения). Может варьироваться возраст аудитории и различать научно-популярную литературу для детей и юношества и для взрослых.

Аналогично в терминах разрабатываемой классификации можно описать тексты, относящиеся к другим сферам функционирования языка – публицистики, художественной литературы, бытового общения. Таким образом, предлагаемый набор параметров служит достаточным основанием для построения типологии текстов и позволяет охарактеризовать любой текст, включаемый в Национальный корпус русского языка.