С ЧЕГО НАЧИНАЕТСЯ ТЕКСТ? ОПЫТ КОРПУСНОГО ИССЛЕДОВАНИЯ [1]
HOW TO START A TEXT? A CORPORA-BASED STUDY

Б. Л. Иомдин (iomdin57@rambler.ru)

Институт русского языка им. В. В. Виноградова РАН

Рассмотрены методы определения местоположения фразы в тексте. Выявлены признаки, характерные для начальных фраз (инверсия, подробные наименования и др.) и фраз, развертывающих или завершающих фрагменты текста (отсылки, эллипсисы и др.). Результаты могут использоваться в системах обработки текстов.

Так, но с чего же начать, какими словами?
(Саша Соколов, Школа для дураков).

Введение

Все бездарные тексты похожи друг на друга, каждый талантливый текст талантлив по-своему. Поэтому сходства и различия в форме текстов привлекали и привлекают внимание множества исследователей. Как структуре построения текстов разных жанров, так и проблематике автоматического определения и категоризации фрагментов текстов посвящена огромная литература по семиотике, литературоведению, фольклористике, этнографии, стилистике, риторике, компьютерной и теоретической лингвистике. К первой, филологической области исследований относятся, прежде всего, работы В. Я. Проппа, Ю. М. Лотмана, Б. А. Успенского, Ю. С. Мартемьянова, Р. Барта. Второй, прикладной проблематикой занимались T. A. van Dijk, P. B. Baxendale, R. Harweg и многие другие (обширные, регулярно обновляемые библиографии можно найти в [14] и [18]). Кроме того, зачинами и концовками текстов занимается стилистика и риторика, см. [9], [15]. Отметим также специальные работы [4], [17], [19].

Специфика данной работы, лежащей на стыке указанных областей науки, состоит в следующем.

1) **Предмет исследования.** Исследование композиции художественного текста, как правило, подразумевает рассмотрение его фрагментов и изучение их формальных и содержательных особенностей в общем контексте. В большинстве прикладных работ упор делается на определение относительной релевантности фрагментов текста в зависимости от частоты встречаемости ключевых слов; см., например, [13]. Таким образом, большинством исследователей анализируются целые тексты. Мы же рассматриваем отдельно взятые фразы без контекста.

2) **Материал исследования**. Внимание литературоведов и фольклористов, изучающих зачины и концовки, прежде всего привлекают специфические тексты: эпос, народные сказки, анекдоты и т. п. (см., например, [1], [10]); отдельная большая литература посвящена рамочной структуре поэтических текстов (см., например, [6]). Круг текстов, которые исследует компьютерная лингвистика, в большинстве случаев обусловлен ее прикладными потребностями: это в первую очередь газетные сообщения, новостные ленты, научно-популярные тексты. В нашей работе рассматривается художественная проза на современном русском языке.

3) **Цель исследования.** Цели, которые ставят перед собой исследователи структуры художественных текстов, весьма разнообразны, однако обычно они не включают в себя анализ отдельных фраз без информации об их относительном расположении в тексте, так как исследователи уже обладают этой информацией (мы отвлекаемся здесь от проблематики изучения древних памятников, сохранившихся лишь фрагментарно). Большинство же работ по компьютерной лингвистике в рассматриваемой области посвящены проблемам автоматической сегментации и реферирования текстов и имеют целью определение информационной нагрузки данной фразы и ее относительной значимости, а также выделение тем (topics) и деление текста на фрагменты (chunks) на этой основе (см., например, [12], [16]). Цель нашей работы – создание методов определения вероятного расположения данной фразы в тексте на основе анализа ее морфологической, синтаксической и семантической структуры.

Материалы и методы

Исследование проводилось на материале корпуса Сектора теоретической семантики Института русского языка им. В. В. Виноградова РАН (в настоящий момент этот корпус содержит более 1500 текстов конца XIX, XX и начала XXI

[1] Работа выполнена при финансовой поддержке грантов Президента РФ для молодых кандидатов наук (№ МК-3992.2004.6) и ведущих научных школ (№ НШ-1576.2003.6), а также гранта РГНФ № 05-04-04190а. Автор хотел бы также выразить благодарность Андрею Санникову за подготовку специальной программы для данного исследования и Льву Иомдину за помощь при проведении мини-эксперимента.

века, принадлежащих к разным жанрам)[2]. Использовались только прозаические тексты (рассказы, повести, романы), в основном написанные в XX веке и не переводные; при этом для бóльшей сбалансированности выборки произведения одного автора не использовались более одного раза. Из соображений места приводим здесь лишь список авторов, чьи произведения были включены в выборку:

Абрамов Ф., Азольский А., Айтматов Ч., Аксенов В., Андреев Л., Антонов С., Астафьев В., Бакин Д., Бакланов Г., Белов В., Булгаков М., Булычев К., Бунин И., Варламов А., Веллер М., Визбор Ю., Войнович В., Волос А., Высоцкий В., Гайдар А., Ганина М., Гаршин В., Гоголь Н., Гончаров И., Гранин Д., Грекова И., Грин А., Гумилев Н., Гуреев М., Дмитриев А., Довлатов С., Достоевский Ф., Екимов Б., Ерофеев В., Етоев А., Зайцев Б., Замятин Е., Зощенко М., Ильф А. и Петров В., Инбер В., Искандер Ф., Кабаков А., Каверин В., Казаков Ю., Козлов С., Конецкий В., Кошко А., Кунин В., Куприн А., Лагин Л., Ларин О., Лермонтов М., Лидин В., Львова М., Максимова С., Маринина А., Нагибин Ю., Незнанский Ф., Носов Н., Одоевский В., Олеша Ю., Павлов Н., Пастернак Б., Паустовский К., Пелевин В., Петрушевская Л., Пикуль В., Пильняк Б., Погодин М., Попов Е., Приставкин А., Пришвин М., Пушкин А., Пьецух В., Семенов Ю., Сенчин Р., Сергиевская И., Соколов С., Солженицын А., Солоухин В., Сомов О., Стругацкие А. и Б., Токарева В., Толстая Т., Толстой Л., Трифонов Ю., Тургенев И., Тэффи, Фадеев А., Филенко А., Хазин В., Хармс Д., Чехов А., Шагинян М., Шаламов В., Шинкарев В., Шмелев И., Шолохов М., Штерн Б., Шукшин И.

Первая и последняя фразы каждого из исследованных произведений были охарактеризованы по следующим пятнадцати параметрам (расположены от более характерных для начальных фраз к более характерным для конечных фраз):

1. Наличие стандартных зачинов (*жили-были, как-то раз, однажды* и т.п.);
2. Наличие конструкции *У X-а есть Y*; ср. *У дяди Кязыма была замечательная скаковая лошадь* (Ф. Искандер, Лошадь дяди Кязыма);
3. Наличие инхоативных форм глаголов и глаголов с инхоативной семантикой; ср. *…Заветрило с ледников, и уже закрадывались по ущельям всюду проникающие резкие ранние сумерки, несущие за собой холодную сизость предстоящей снежной ночи* (Ч. Айтматов, Плаха); *День для нее начинался с неистового убеждения себя в том, что у нее есть сын…* (Д. Бакин, Стражник лжи);
4. Указание времени действия;
5. Указание рода занятий персонажей;
6. Указание места действия (топонимы);
7. Инверсия (сказуемое предшествует подлежащему); ср. *Первыми увидели немцев разведчики* (Г. Бакланов, Мертвые сраму не имут)[3];
8. Вопросительное предложение; ср. *Как быть с вещами, которых мы не помним?* (В. Инбер, Смерть луны);
9. Указание даты[4];
10. Имена персонажей;
11. Наличие эллипсиса; ср. *И услышал пение тетивы* (А. Филенко, Шествие динозавров);
12. Наличие вводных слов; ср. *Впрочем, я был, наверное, слишком избалован для того, чтобы есть суп из одного только зеленого лука* (А. Варламов, Дом в деревне);
13. Наличие местоимений третьего лица;
14. Наличие слов, отсылающих к предшествующему контексту (ср. *другой, тот же, следующий, теперь*);
15. Сочинительный союз в начале предложения.

Результаты

Результаты подсчетов суммированы в Таблице 1.

Наиболее релевантными оказались параметры 1–7 для начальных фраз и 13–15 для конечных. Отметим, что 1 и 2 характерны исключительно для начальных, а 14 и 15 – исключительно для конечных фраз (во всяком случае, на всем исследованном материале не встретилось ни одного контрпримера).

Таким образом, по нашим (весьма предварительным) данным первые фразы произведений обладают бóльшим числом отличительных признаков, чем последние. Для проверки этой гипотезы и верификации предложенного метода подсчета был проведен следующий мини-эксперимент.

Для 30 фраз, десять из которых являются начальными фразами произведений, десять – конечными, и еще десять были взяты из середины, были проведены подсчеты для определения вероятных начальных фраз (то есть таких, у которых отношение F/N и/или FN/LN больше 1), следующим образом. Общий удельный вес параметров, говорящих в пользу того, что данная фраза

[2] См. подробно о составе этого корпуса в списке источников Нового объяснительного словаря синонимов русского языка [2: LV–LXVIII].

[3] То, что инверсия типична для текстов нарративного жанра, в особенности анекдотов, отмечалось, в частности, в работах [10: 527] и [11: 181].

[4] Неточные указания времени и дат (ср. *Не помню, в каком году, но где-то далеко после войны я плыл на новом пароходе вниз по Енисею* (В. Астафьев, Без приюта)), учитывая ограниченные возможности компьютерного анализа текстов при потенциальной автоматизации предлагаемой процедуры, не принимались во внимание, хотя такие указания встречаются достаточно часто. Ср.: *Облачным, но светлым днем, в исходе четвертого часа, первого апреля 192… года (иностранный критик заметил как-то, что хотя многие романы, все немецкие например, начинаются с даты, только русские авторы – в силу оригинальной честности нашей литературы – не договаривают единиц)…* (В. Набоков, Дар).

Параметр	First, %	Last, %	FirstN, %	LastN, %	F/L, %	FN/LN, %
1	4	0	4	0	∞	∞
2	1	0	1	0	∞	∞
3	7	1	7	1	7,00	7,00
4	7	1	7	1	7,00	7,00
5	20	2	11	2	10,00	5,50
6	45	9	29	7	5,00	4,14
7	16	5	16	5	3,20	3,20
8	3	1	3	1	3,00	3,00
9	5	2	4	2	2,50	2,00
10	44	27	37	20	1,63	1,85
11	1	2	1	2	0,50	0,50
12	1	2	1	2	0,50	0,50
13	21	68	15	36	0,31	0,42
14	0	12	0	9	0,00	0,00
15	0	19	0	19	0,00	0,00

Таблица 1. *Характеристики первых и последних фраз рассмотренных текстов. В столбцах «First» и «Last» указано общее число элементов фраз (соответственно, начальных и конечных), содержащих данный параметр; в столбцах «FirstN» и «LastN» приведены цифры при условии, что каждый из параметров принимает лишь значения 0 или 1*[5]*. В столбцах «F/L» и «FN/LN» соответственно приведены отношения суммарного значения данного параметра в начальных фразах к его суммарному значению в конечных фразах при обоих способах подсчета.*

Фраза	Статус	P_F/P_L	P_FN/P_LN
1	F	2,48	1,84
2	L	0,31	0,22
3	M	0	0
4	F	7,00	7,00
5	L	0,38	0,55
6	F	0,50	0,50
7	F	1,63	1,37
8	M	0	0
9	L	0,00	0,00
10	F	3,20	3,20
11	L	1,63	1,37
12	L	0	0
13	F	0	0
14	M	3,20	3,20
15	L	2,91	2,16

Фраза	Статус	P_F/P_L	P_FN/P_LN
16	L	0	0
17	M	0,31	0,22
18	F	5,20	3,60
19	M	1,36	1,20
20	F	1,13	0,95
21	M	0	0
22	M	0,31	0,22
23	L	0	0
24	L	1,87	1,41
25	M	0,31	0,22
26	L	0,28	0,19
27	M	0	0
28	M	0	0
29	F	2,50	2,00
30	F	3,83	3,83

Таблица 2. *Оценка вероятного расположения фраз в тексте по предлагаемому методу. В столбце «Статус» указано действительное расположение фразы в тексте (F – начальная, L – конечная, M – взятая из середины).*

является начальной – $\mathbf{P_F}$ (и, соответственно, конечной – $\mathbf{P_L}$), складывается из значений по всем 15 параметрам, домноженных на соответствующие суммарные значения F/L[6] для начальных (соответственно, конечных) фраз, приведенные в Таблице 1. По итоговым цифрам для каждой фразы рассчитывается отношение $\mathbf{P_F/P_L}$. Вероятно начальными признаются те фразы, у которых это отношение оказывается больше 1 – и тем вероятнее, чем больше. Результаты подсчетов представлены в Таблице 2.

[5] Например, для фразы *А за ними стояла и сильнее их влекла его к себе – еще другая, исполинская фигура, другая великая бабушка – Россия* (И. Гончаров, Обрыв) во втором столбце учтены значения «3» по параметру 13 (местоимения третьего лица) и «2» по параметру 14 (слова типа *другой*), а в четвертом столбце – лишь «1» по обоим параметрам.

[6] Или FN/LN, при втором способе подсчета, результаты которого обозначены $\mathbf{P_FN}$ и $\mathbf{P_LN}$, соответственно.

Итак, из 10 начальных фраз 8 угаданы верно, в одной (№13) не удалось выделить признаков, характерных для начальных или конечных фраз (суммарное значение по всем параметрам равно 0) и одна (№6) признана скорее конечной. Вот эти фразы: (№13) *Мысли о недавнем гулянье становились все мрачнее, навевали тоску и унынье* (Н. Гладышев, Антонов колодец); (№6) *Да, этот год не то, что прошлый* (В. Каплан, Усатый-полосатый).

Кроме того, еще 5 фраз были ошибочно признаны вероятно начальными. Вот эти фразы: (№11) *Лорд Рокстон посмотрел на меня и молча протянул мне свою крепкую, загорелую руку* (А. Конан-Дойль, Затерянный мир. Пер. Н. Волжиной); (№14) *На камне появилась ящерица* (Ю. Олеша, Любовь); (№15) *На экране возник зал заседаний в Кремле и лицо председателя Верховного Совета Анатолия Лукьянова* (А. Житинский, Параллельный мальчик); (№19) *Неподвижно, еще не в силах сообразить происшедшее, стояла гибкая Зара, прислонясь к узорчатой стене* (Н. Гумилев, Принцесса Зара); (№24) *По утверждению обезумевшего поручика, на третий день, вечером, голова Геро крикнула страшным, глубоким и как бы мужским голосом* (М. Павич, Внутренняя сторона ветра. Пер. Л. Савельевой). Очевидно, что за бóльшую часть ошибок ответственны имена персонажей, встретившиеся в этих фразах.

Во второй части мини-эксперимента те же 30 фраз были предъявлены испытуемым, которые должны были определить, являются ли данные фразы скорее начальными, скорее конечными, или они взяты из середины. Местоположение фразы №6 угадал лишь один испытуемый, фразы №13 – ни один. Что касается пяти фраз, ошибочно признанных начальными в ходе первой части эксперимента, то при определении вероятного расположения в тексте каждой из них испытуемые также допустили ошибки (в частности, фразу №15 последней не признал никто). В целом же испытуемые правильно определили в среднем 5,5 из 10 начальных фраз.

Таким образом, мини-эксперимент показывает, что результаты подсчетов по предлагаемому методу по крайней мере не хуже результатов, полученных на основе интуиций носителей языка.

Возможные области применения

Идея настоящей работы может показаться достаточно искусственной. Однако можно сформулировать задачи из различных областей науки (как прикладные узкоспециальные, так и представляющие теоретический интерес), где ее результаты могли бы пригодиться.

1) Автоматическое структурирование корпусов текстов. Корпусная лингвистика сейчас переживает бурный расцвет, а количество привлекаемых ею текстов растет лавинообразно. Нередки ситуации, когда исследователю приходится вручную разделять и структурировать огромные массивы текстов, в частности найденных на просторах Интернета, или, напротив, соединять в единое целое произведения, разбитые на множество файлов для удобства хранения или поиска. Безусловно, в любом корпусе, предназначенном для серьезных исследований, каждый текст должен быть тщательно выверен, однако для первичной «грубой» обработки метод автоматического определения начальных и заключительных фраз может оказаться удобным.

2) Оценка качества подготовленного корпуса текстов. Когда корпус состоит из многих тысяч файлов, количество которых к тому же постоянно увеличивается, на их тщательную обработку может не хватать времени. Процедура, позволяющая хотя бы частично автоматизировать процесс выявления неполных, оборванных текстов без начала или конца, могла бы внести свой вклад в повышение качества корпусов.

3) Систематизация документации. Специфика существования текстов в электронную эпоху не всегда подразумевает, что документы хранятся в четко структурированном и классифицированном виде, как на локальных компьютерах, так и в сетях. Процедура, подобная предлагаемой, могла бы войти в качестве одного из модулей в алгоритм структурирования и систематизации большого количества текстов разной степени оформленности и подготовленности.

4) Изучение развития стиля литературных произведений. Со временем представления авторов о том, как именно должны начинаться и заканчиваться литературные произведения тех или иных жанров, меняются. Процитируем Н. Д. Арутюнову: “А. Белый один из первых заметил изменения в формах начала и конца современных ему художественных текстов. Его внимание привлекла первая фраза романа С. Пшибышевского «Homo sapiens» («*Фальк вскочил окончательно взбешенный*»), а также аналогичные начала других произведений этого автора. Вот как он охарактеризовал такой тип начала (*бросок с места в карьер*) <…>: «Одна фраза и в ней водораздел двух стихий, отделяющий творчество великих писателей середины XIX столетия от писателей конца века; одна фраза – а мы уже чувствуем: что-то произошло. У писателей доброго старого времени <…> после долгих пояснений автора появлялся герой со своими поступками, встреченный нами, как старый знакомец. А вот у Пшибышевского прямо из мрака неизвестности на сцену выскакивает какой-то Фальк и начинает перед нами судорожно беситься»” [3: 16][7].

[7] Ср. также характерную цитату из современной научной фантастики, отражающую закономерное развитие тенденции, отмеченной А. Белым: “После заголовка самое важное – первая фраза. Она должна быть как удар гонга, как отдернутый занавес, как вспышка магния в

Собранный материал вкупе с выделенными языковыми характеристиками первых и последних фраз мог бы дать богатую почву литературоведам и лингвистам для выявления особенностей текстов той или иной эпохи.

5) Сравнение различных типов дискурса. Еще шире, сравнительный анализ по предлагаемому методу дискурса различных типов (как письменного, так и устного) также представляет интерес (о результатах одного из исследований устной речи на сходную тему на материале английского, испанского и шведского языков см. работу [20]; см. также [8]).

Заключение

Нельзя не сознавать, что литература – свободное пространство, в котором каждый автор волен поступать так, как он считает нужным. Поэтому, безусловно, первые фразы произведений могут обладать всеми признаками последних (ср. хотя бы знаменитое начало рассказа В. Набокова «Круг»: *Во-вторых: потому что в нем разыгралась бешеная тоска по России*), и наоборот; ср. последнюю фразу *Приближался девятьсот пятый год* (В. Инбер, Смерть луны). Однако при статистически значимой выборке такие яркие исключения, как представляется, должны нивелироваться. Тщательный анализ такой выборки, существенное расширение материала, усовершенствование предлагаемого метода, несомненно, необходимы для полноценного выполнения любой из сформулированных выше задач. Но наше исследование отнюдь не претендует на законченность, а имеет целью лишь обозначить возможное направление работы. В условиях становления новых форм бытования текстов и возникновения целого пласта новых задач, связанных с этим, можно использовать накопленный опыт лингвистов, литературоведов, исследователей структуры текста для их решения и продемонстрировать, что даже такая небольшая модель может давать результаты.

Если попытаться осмыслить те наборы параметров, которые оказались наиболее релевантными для определения расположения фразы, то окажется, что для начальных фраз текстов характерна бóльшая информационная нагруженность: в них вводятся персонажи, указывается их род занятий, отмечается время и место действия. Напротив, заключительные фразы отсылают к предшествующему контексту, без которого они подчас непонятны: в них чаще используется анафора, эллиптические конструкции. Это отражает естественную структуру построения текста: обычно к его концу все персонажи уже введены, их действия описаны, и о возможном развитии событий предоставляется судить читателю. Как писал Ю. С. Мартемьянов, при понимании любого текста его читателю необходимо “сопоставить уже выявленным из текста голым фактам ситуации – их место в некоторой более широкой картине действительности, указать на связи их с определенной системой более глубоких представлений о той же действительности” [7: 125–126]. Стало быть, и по прочтении заключительной фразы текст продолжает жить в сознании читателя, и не кончается строка.

Список литературы

1. Антонов Д. Концовки волшебных сказок: попытка прочтения // Фольклор и постфольклор: структура, типология, семиотика. http://www.ruthenia.ru/folklore/antonov1.htm.

2. Апресян В.Ю., Апресян Ю.Д., Бабаева Е.Э., Богуславская О.Ю., Галактионова И.В., Гловинская М.Я., Григорьева С.А., Иомдин Б.Л., Крылова Т.В., Левонтина И.Б., Птенцова А.В., Санников А.В., Урысон Е.В. Новый объяснительный словарь синонимов русского языка. 2-е издание, испр. и доп. Под общим руководством акад. Ю. Д. Апресяна. М.–Вена, 2004.

3. Арутюнова Н.Д. В целом о целом. Время и пространство в концептуализации действительности // Логический анализ языка. Семантика начала и конца. М., 2002. С. 3–18.

4. Веллер М. Технология рассказа. М., 1989.

5. Казакевич О.А. Начало и конец в структуре фольклорных текстов северных селькупов // Логический анализ языка. Семантика начала и конца. М., 2002. С. 542–550.

6. Каргашин И.А. Начало и конец лирического текста // Логический анализ языка. Семантика начала и конца. М., 2002. С. 426–435.

7. Мартемьянов Ю.С. Заметки о строении ситуации и форме ее описания // *Машинный перевод и прикладная лингвистика.* 1964. Вып. 8. С. 125–149.

8. Новикова Н.С., Серова Л.К., Щербакова О.М., Попова М.Т. Стереотипы естественной речи и практика речевой коммуникации (прикладной аспект) // Теория коммуникации и прикладная коммуникация. Вестник Российской коммуникативной ассоциации, выпуск 1. Ростов н/Д, 2002. С. 98–108.

9. Солганик Г.Я. Стилистика текста: Учебное пособие. М., 1997.

10. Шмелева Е.Я. Начало и конец русского анекдота // Логический анализ языка. Семантика начала и конца. М., 2002. С. 523–528.

11. Янко Т.Е. Коммуникативные стратегии русской речи. М., 2001.

12. Allan J., ed. Topic Detection and Tracking: Event Based Information Retrieval. Kluwer Academic Press, 2002.

темноте. Нужно, чтобы читатель вошел в книгу, как выходят с чердака на крышу, и увидел бы всю историю до самого горизонта” (Г. Гуревич, Только обгон).

13. Chen Kuang-hua, Chen Hsin-Hsi. A Corpus-Based Approach to Text Partition // Proceedings of the Workshop of Recent Advances in Natural Language Processing, 1995. Pp. 152–161.

14. Computational Linguistics and Information Retrieval. http://tangra.si.umich.edu/clair/home/summ-bib.html

15. Crew L. Rhetorical Beginnings, Professional and Amateur // *College Composition and Communication* 38.3 (1987). Pp. 346–350.

16. Fukumoto F., Suzukit Y., Fukumoto J. An Automatic Extraction of Key Paragraphs Based on Context Dependency. Applied Natural Language Processing 1997. Pp. 291–298.

17. Harweg R. Beginning a Text // *Discourse Processes*, 3 (Dec. 1980). Pp. 313–326.

18. Mallett D. Text summarization: an annotated bibliography. http://www.cs.ualberta.ca/~mallett/biblio.pdf

19. Tichy H.J. Effective Writing for Engineers, Managers, Scientists. John Wiley & Sons, 1966.

20. Tolchinsky L., Johansson V., Zamora A. Text openings and closings in writing and speech: Autonomy and differentiation // Written Language & Literacy, 2002, vol. 5, No. 2. Pp. 219–252.