АКТУАЛЬНІ ПРОБЛЕМИ ЛІНГВІСТИКИ

УДК 811.161.1’38

Е.И. ПАНЧЕНКО,
доктор филологических наук,
заведующая кафедрой лингвистической подготовки иностранцев
Днепропетровского национального университета им. Олеся Гончара

ОБЪЕМ ТЕКСТА КАК ПОКАЗАТЕЛЬ ЕГО СЖАТОСТИ

В статье рассматривается проблема статистических параметров особого вида текстового образования – сжатого текста.

Ключевые слова: текст, сжатый текст, объем, графема, слово.

В данной статье в общем виде ставится проблема исследования одного из аспектов сжатого текста – его объема. Актуальность данного исследования обусловлена значительной ролью сжатых текстов во многих сферах деятельности человека (реферирование, аннотирование, сообщение новостей, SMS-переписка и др.).

Под сжатым текстом мы понимаем сообщение, объективированное подобно любому иному тексту в письменной форме, построенное путем сокращения полного текста либо созданное как изначально краткое, предназначенное при необходимости для дальнейшего развертывания в более объемный текст. Сжатый текст имеет повышенную информативную насыщенность по сравнению с первичным полным вариантом, что достигается благодаря разнообразным обязательным и факультативным средствам всех уровней языка.

Объем сжатого текста (СТ) можно считать одним из важных его признаков, хотя для ряда текстов (конспект, тезисы) этот признак, по-видимому, является факультативным. Такой признак СТ, как его объем, на наш взгляд, непосредственно связан с его лингвистическими характеристиками, так как, по нашим наблюдениям, существует определенная корреляция объема текста как с планом содержания (характер и тематика СТ), так и с планом выражения (в частности, выбор грамматических способов передачи необходимой информации). Исследование краткости и компрессии текста ведут Б.П. Дюндик, И.И. Инфантова, К.М. Сухенко и многие другие авторы, однако проблема, связанная с влиянием объема текста на процесс и результат его сжатия еще остается нерешенной.

Целью данной статьи является установление соотношения полного текста и его сжатого варианта на уровне графем, лексем и предложений.

По нашим наблюдениям, существуют две основные разновидности таких текстов: первичный сжатый текст, формирующийся до создания полного варианта, и вторичный сжатый текст, являющийся результатом переработки соответствующего развернутого текста. Первичный текст может быть потенциально развертываемым (словарная статья, телеграмма) или необходимо развертываемым (проспект издания, план сочинения). Потенциально развертываемый текст может быть независимым, то есть понятным без контекста и ситуации, ситуативно или контекстуально зависимым или опираться на экстралингвистическую основу (подпись к рисунку). Необходимо развертываемый текст может быть подготовлен для публикации и существовать самостоятельно (тезисы научного сообщения) или

быть вспомогательным звеном при построении полного варианта (план выступления). В свою очередь, вторичный сжатый текст может существовать как неотъемлемая составная часть полного варианта либо как отдельное речевое произведение.

Наиболее актуальной является проблема объема документа и степени его сжатия для таких распространенных жанров СТ, как реферат и аннотация того или иного издания, чаще всего научного. Как указывается в лингвистической литературе, научный текст адресован узкому кругу специалистов и рассчитан на наличие общего опыта, знаний, интересов и целей. Поскольку коммуникация в науке есть непрерывный процесс, отправитель сообщения описывает не все детали денотативной ситуации, а лишь новую информацию и выявляет связи этого нового с уже известным, что осуществляется в форме реферата и аннотации.

Существуют определенные стандарты, касающиеся объема указанных жанров, хотя многообразие назначения и видов рефератов и аннотаций иногда заставляет их составителей отходить от таких стандартов. Согласно ГОСТу, объем реферата зависит от объема реферируемого издания и может составлять от 500 печатных знаков и кратких сообщений до 2500 знаков для документов большого объема. Средний объем аннотации составляет около 600 печатных знаков.

Таким образом, теоретически определены верхний и нижний пределы указанных документов, и мы поставим перед собой вопрос: как практически соотносятся объем первичного и вторичного документов?

Как неоднократно подчеркивалось в лингвистических работах, исследование различных аспектов того или иного текста целесообразно проводить с учетом количественных измерений. Количественные подсчеты не являются основной целью нашего исследования, и мы не претендуем на их полноту и завершенность, считая, что для специальных исследований именно в этом направлении нужны особые методы и несравненно большее количество материала, чем то, которое отобрано нами для получения некоторых количественных иллюстраций. Поэтому основное внимание в нашем исследовании мы уделяем не абсолютным, а относительным количественным показателям, касающимся объема документа.

Для наших количественных исследований мы сосредоточим внимание на соотношении таких речевых произведений, как первичный текст (книга, статья, диссертация) и относящиеся к нему вторичные документы – реферат и аннотация. Первичный текст является исходным материалом, который подвергается процессу сжатия при необходимости представить его в той или иной минимизированной форме. В первую очередь в таком случае происходят изменения в семантическом и семантико-логическом плане. Тем не менее, возможно, существует определенная закономерность между объемом первоисточника и объемом того текста, который возникает при его сокращении. Для данного исследования мы произвели соответствующий отбор первоисточников и указанных СТ, сформировав выборку в количестве 100 текстов-первоисточников, 100 рефератов, 100 аннотаций. В данную выборку вошли речевые произведения с различной содержательной направленностью: работы по техническим дисциплинам (машиностроение, энергетика и другие), а также теоретические произведения в области языкознания и литературоведения, опубликованные в реферативных журналах.

Для получения усредненных данных тексты были сгруппированы в 10 серий по 10 единиц в каждой и по указанным сериям были проведены определенные количественные подсчеты. Проведение количественных подсчетов имело целью решение следующих задач: определение абсолютного и относительного соотношения объемов первичного и вторичного документов по количеству графем; определение такого же соотношения по количеству слов и предложений; определение степени сжатия и ее зависимость от объема первоисточника; исследование степени использования новых слов в указанных жанрах СТ.

Необходимо дать следующие пояснения к выборкам, результаты анализа которых представлены в табл. 1–3. В выборки 1–8 вошли тексты, представляющие собой научные работы по техническим дисциплинам; рефераты и аннотации к ним соответствуют названным выше требованиям к их построению. Выборки 9 и 10 включают тексты по филологическим проблемам (языкознанию и литературоведению). Рефераты к ним представлены в соответствующих реферативных журналах. Эти рефераты имеют значительно больший объем по сравнению с техническими стандартными текстами и, как нам представляется,

являются рефератами по названию, а по сути приближаются к жанру рецензии и не имеют четко определенных объемных рамок.

Приведем примеры текстов рефератов и аннотаций, которые послужили объектом нашего исследования.

Аннотация. В книге рассмотрена классификация машин для подачи жидкостей и газов, изложены кратко основы теории этих машин, освещены вопросы ориентировочных аэродинамического и гидродинамического расчетов некоторых типов насосов и компрессоров. В книге рассмотрены конструкции машин и их элементов. Уделено внимание некоторым новым оригинальным конструкциям машин.

Книга предназначена для ИТР, занимающихся исследованием, проектированием, изготовлением и эксплуатацией насосов. Она может быть также полезна студентам машиностроительных специальностей.

Реферат. Сообщается о классификации насосов, их особенностях и полях применения. Далее раскрываются типовые х-ки всех основных групп лопастных и объемных насосов (в том числе и с магнитным приводом), включая также и зависимости подачи и функции времени (для объемных машин). Рассматриваются конструктивные схемы и особенности каждой типовой конструкции. Наибольшее внимание уделяется новым и перспективным конструкциям – насосам с магнитным приводом, дозировочным насосам повышенной точности дозирования однофазных и двухфазных жидкостей, а также центробежным насосам высокого давления. Сформулированы основные требования к современному насосостроению: высокий КПД, надежность, автоматическое регулирование процесса, герметизация, применение пластмасс.

Результаты количественных подсчетов по приведенным выше текстам суммированы в последующих таблицах. В табл. 1 представлено соотношение объема первичного и вторичного документов (реферата и аннотации), выраженного в графемах. Подсчет количества графем в полном варианте (первичном тексте) был сделан приблизительно, подсчитывалось их среднее количество на нескольких страницах и затем проводился общий подсчет по всему объему текста. Как видно из приведенной таблицы, в среднем реферат и аннотация составляют 1–1,8% объема текста-первоисточника. Если сравнить полученные данные с требованиями стандарта к реферату, аннотации, то можно сделать вывод о том, что объем аннотации несколько меньше, чем это отмечается в стандарте, тогда как объем реферата примерно соответствует стандартным требованиям, однако различия в объеме рефератов являются более заметными.

Таблица 1

Объем первичного и вторичного текстов в графемах

Вид текста № выборки	Документ	Реферат	% к 2	Аннотация	% к 2
1	2	3	4	5	6
1	8250	91	1,1	78	0,95
2	33000	370	1,1	363	1,1
3	22000	105	0,5	122	0,6
4	24750	644	2,6	511	2,1
5	13750	595	4,3	364	2,65
6	30250	966	3,2	501	1,7
7	16500	294	1,8	237	1,4
8	27500	182	0,7	206	0, 75
9	657250	6201	0,9	434	0,07
10	189750	3475	1,8	392	0,2
Среднее	102300	1292	1,8	320,8	1,14

Рассмотрим далее полученные нами подсчеты, касающиеся объема первичного и вторичного документов, выраженного в лексических единицах и предложениях. Эти под-

счеты представлены в табл. 2 и 3. Подсчет количества лексических единиц в полном варианте (первичном тексте) был сделан приблизительно, то есть нами подсчитывалось среднее количество слов на нескольких страницах (около десяти), а затем проводился общий подсчет по всему объему анализируемого текста.

Таблица 2

Объем первичного и вторичного текстов в словах

Вид текста № выборки	Документ	Реферат	% к 2	Аннотация	% к 2
1	2	3	4	5	6
1	900	13	1,4	11	1,2
2	3600	53	1,5	50	1,4
3	2400	15	0,6	18	0,8
4	1500	85	5,7	52	3,5
5	3300	138	4,2	70	2,1
6	1800	42	2,3	38	2,1
7	3000	26	0,9	33	1,1
8	3600	210	5,8	78	2,2
9	71700	1120	1,6	62	0,09
10	20700	462	2,2	56	0,27
Среднее	11250	216	2,6	46,8	1,48

Таблица 3

Объем первичного и вторичного текстов в предложениях

Вид текста № выборки	Документ	Реферат	% к 2	Аннотация	% к 2
1	2	3	4	5	6
1	48	3	6,3	3	6,3
2	204	4	1,96	4	1,96
3	128	2	1,6	2	1,6
4	144	4	2,8	3	2,1
5	85	3	3,5	4	4,7
6	188	5	2,7	4	2,1
7	96	3	3,1	2	2,1
8	170	2	1,2	4	2,4
9	3824	39	1,02	4	0,1
10	1104	19	1,7	3	0,3
Среднее	599	8,4	2,6	3,3	2,4

В результате анализа количественных данных, представленных в табл. 1–3, можно сделать следующие выводы. Количество слов и предложений в рефератах и аннотациях произведений по техническим дисциплинам сопоставимо между собой, в некоторых случаях рефераты несколько длиннее аннотаций, в других случаях – наоборот, хотя в основном аннотации незначительно короче. В выборках 9 и 10 по филологическим проблемам рефераты значительно длиннее аннотаций (в среднем в 10 раз), тогда как аннотации к указанным текстам соответствуют общему стандарту и по объему практически не отличаются от имеющихся аннотаций к техническим произведениям.

В определенной мере изменение объема сжатых текстов, выраженного в количестве слов и предложений, зависит не только от объема полного текста, но и от того издания, в котором рефераты опубликованы, иными словами, от личности составителя и целей, которые он ставит перед собой, то есть реферат несет отпечаток субъекта творческого процесса. Это заметно в выборках 3, 7 и 8, когда объемы текстов в выборках 3 и 7 (РЖ «Машиностроение») оказываются несколько меньше средних, а в выборке 8 (РЖ «Насосостроение») несколько больше.

Исходя из соотношения размеров вторичных документов (реферата и аннотации) и первичного документа, можно также отметить, что в среднем реферат и аннотация по количеству слов и предложений составляют около 2% объема первоисточника, то есть формально текст сворачивается примерно в 1/50 объема, тогда как в лингвистических исследованиях отмечается, что оптимальная цифра составляет 1/8.

Количественные данные по различным видам сжатого текста позволяют предположить, что объем сжимаемого текста косвенным образом влияет на объем текста сжатого, хотя представляется затруднительным определить их прямую зависимость. Как свидетельствуют наши исследования, при объеме документа более 20 страниц, то есть около 6000 слов, объем реферата перестает увеличиваться. Объем же аннотации практически не зависит от объема аннотируемого произведения.

Еще одним направлением количественного анализа сжатых текстов является исследование разнообразия их лексического состава. Процентное соотношение разных слов на фиксированном отрезке текста было проанализировано нами по нескольким методикам. В первом случае в силу того, что, как правило, сжатые тексты представляют собой достаточно короткое речевое произведение, мы осуществили 100 выборок текстов рефератов и аннотаций длиной по 100 слов в каждой выборке. В соответствующих текстах первоисточников также были отобраны отрезки равного объема. В каждой выборке было определено процентное содержание новых слов по сравнению с предыдущей. Несколько иное в количественном плане второе исследование процентного соотношения разных слов в документе – реферате – аннотации было проведено на текстах, посвященных проблемам: 1) насосостроения; 2) двигателестроения; 3) языкознания. Была проанализирована выборка из 1000 слов первоисточника и такой же объем аннотаций и рефератов по указанным проблемам. Анализ полученных данных свидетельствует о том, что количество новых слов в таком виде выборки выше в реферате и аннотации по сравнению с текстом первоисточника; аннотации можно считать несколько более разнообразными по составу, чем реферат (на 2–3%). Различие в степени разнообразия слов в первом и втором случае, по-видимому, связано с тем, что в 10 выборках анализу подвергались разные тексты-первоисточники, тогда как во втором случае аннотации и рефераты представляли собой сжатый вариант одного и того же произведения, а 1000 слов были отобраны из одного первичного текста. Иными словами, степень разнообразия лексических единиц в определенной мере связана с объемом СТ.

Было также исследовано количество разных слов в полном тексте – аннотации – реферате методом обратного наложения. При этом мы проанализировали 50 аннотаций и 50 рефератов по техническим вопросам и соответствующее их объему количество слов в первоисточнике, отобранных в его произвольно взятой части. При этом процент разных слов в документе составил в среднем 79%; в аннотации – 83%; в реферате – 84%.

Таким образом, можно отметить, что объем текста реферата имеет косвенную зависимость от объема текста-первоисточника, тогда как объем аннотации является более стандартизованным и определяется внутренними законами построения этого жанра. Этот объем наполняется достаточно разнообразными по составу единицами, степень разнообразия которых в реферате и аннотации примерно одинакова.

У статті розглядається проблема статистичних параметрів особливої текстової побудови – стислого тексту.

Ключові слова: текст, стислий текст, обсяг, графема, слово.

The article deals with the problem of statistic parameters of compressed texts as a special textual construction.

Key words: text, compressed text, volume, grapheme, word.

Надійшло до редакції 8.02.2011.