

РАБОЧАЯ ИНСТРУКЦИЯ

ПОДГОТОВКА ТЕКСТОВЫХ МАТЕРИАЛОВ

1. ВВЕДЕНИЕ

1.1 Определение понятия «Подготовка текстовых материалов»

Подготовка текстовых материалов — получение требуемых для дальнейшей работы текстовых материалов в электронной форме в виде документов Microsoft Word. Подготовка текстовых материалов может включать в себя следующие действия:

1. Распознавание текста (в том числе таблиц с числовыми данными), имеющегося в виде отсканированной бумажной копии.
2. Распознавание текста (в том числе таблиц с числовыми данными), имеющегося в виде файла формата PDF.
3. Извлечение (копирование) текста из файлов, не являющихся файлами формата Microsoft Word, но содержащих текст в электронном (не растровом) виде (например, PDF, Microsoft PowerPoint, и проч.).
4. Обработка файлов Microsoft Word, полученных после распознавания или извлечения текста из файлов.
5. Набор текста (в том числе таблиц с числовыми данными).
6. Набор математических формул.
7. Подготовка файлов формата Microsoft Office к переводу: обеспечение редактируемости текста.

2. ОСНОВНЫЕ ПОНЯТИЯ И ТРЕБОВАНИЯ

2.1 Описание

Исходными материалами для начала подготовки текстов могут быть:

1. Материалы заказчика в бумажном виде (копия оригинала либо распечатка файла).
2. Материалы заказчика в электронном виде (для распознавания или извлечения текста) — файлы оригинала, либо отсканированные (для распознавания либо извлечения текста копированием).
3. Дополнительные требования к подготовке текста, указанные менеджером проекта.

Результатом подготовки текста является файл(-ы) в формате Microsoft Word и иных форматах Microsoft Office, содержащий текст, таблицы с числовыми данными и уравнения (при необходимости), удовлетворяющие изложенным в данной инструкции требованиям. На основе одного оригинала может создаваться более одного файла, например, (1) файл с текстом для дальнейшего использования в системах TranslationMemory, (2) файл, содержащий таблицы с числовыми данными и (3) файл, содержащий уравнения для дальнейшего их использования в процессе верстки. Если специально не указано, то результатом должен быть только один файл..

Текст в результирующем файле не должен содержать пропусков и/или символов, отсутствующих в оригинале, если иное специально не указано в задании.

2.1.1 Набор текста (таблицы с числовыми данными, уравнения)

Набор текста (в том числе таблиц с числовыми данными и уравнений) производится при отсутствии возможности получить текст каким-либо другим способом, например, при плохом качестве бумажной копии оригинала, когда распознавание невозможно. Текст (в том числе таблицы) набирается с клавиатуры в программе Microsoft Word на требуемом языке с соблюдением орфографии оригинала. Стиль оформления текста соблюдать не следует, за исключением сохранения выделения элементов текста (курсив, полужирное начертание, подчеркивание и т. п.). Уравнения набираются только при помощи штатной функции программы Microsoft Word (Вставка – Формула), при этом следует соблюдать начертание шрифта оригинала (рубленный либо с засечками) и размер шрифта. Таблицы следует набирать с сохранением структуры таблиц оригинала и с сохранением выделения элементов текста. При наборе таблиц и уравнений следует учитывать, на какой язык будет выполняться перевод, и соответствующим образом выполнять расстановку знаков десятичной запятой (точки) и разделителей разрядов тысяч (при необходимости).

2.1.2 Копирование текста

Текст может быть скопирован из файла оригинала в файл Microsoft Word в случае, если это необходимо и является более продуктивным, чем распознавание.

2.1.3 Распознавание текста

Для распознавания текста могут применяться любые программы, если они обеспечивают выполнение требований задания на распознавание. Рекомендуется использовать программу ABBYY FineReader, так как по сравнению с другими аналогичными программами, а также по сравнению с программой Microsoft Word, она имеет наиболее полный список словарей для проверки орфографии распознанного текста.

При необходимости возможно выполнение следующих подготовительных действий:

1. Сканирование бумажной копии целесообразно выполнять отдельным действием при помощи программ, поставляемых со сканерами, в соответствии с инструкцией по сканированию, а не как операцию, выполняемую в программе распознавания.

2. При необходимости файлы, полученные в результате сканирования, могут быть дополнительно обработаны в программах обработки растровых графических изображений (например, для повышения контрастности текста и четкости (резкости) текста при работе с оригиналом плохого качества, для разворота изображения так, чтобы строки текста располагались строго горизонтально, и т. п.) для снижения количества возможных ошибок распознавания.

3. При распознавании растровых файлов с текстом, размер шрифта которого равен 6 pt и менее (если текст такого размера занимает не менее ~20%), допускается (рекомендуется) предварительно увеличить разрешение растрового файла до величины не менее 500 dpi.

4. При наличии в графических файлах фона под текстом, затрудняющего распознавание, допускается (рекомендуется) выполнить предварительную обработку таких файлов для устранения влияния фона.

5. При использовании программы ABBYY FineReader следует:

5.1. Разметку блоков распознавания рекомендуется выполнить вручную на каждой странице, при распознавании таблиц следует удостовериться в правильном определении структуры таблицы программой, при необходимости исправить.

5.2 В качестве параметров сохранения распознанного файла (ABBYY FineReader версии 12 и более новые) выбрать для файлов DOCX «Оформление документа – Простой текст, Сохранять полужирный шрифт, курсив и подчеркивание», параметр «Сохранять картинки» должен быть отключен. Использование параметра «Сохранять полное оформление документа» не допускается. «Шрифт, используемый в режиме Простой текст – Arial». Для более старых версий программы следует выбирать наиболее «простой» режим сохранения, позволяющий сохранить текстовые выделения оригинала.

5.3 Таблицы с числовыми данными следует проверить после распознавания в программе ABBYY FineReader, обращая внимание на маркировку «неуверенно распознанных символов».

5.4 Следует проверить орфографию распознанного текста (не числовых данных) средствами программы ABBYY FineReader, при этом следует включить параметр проверки «Останавливаться на несловарных словах» и отключить все остальные.

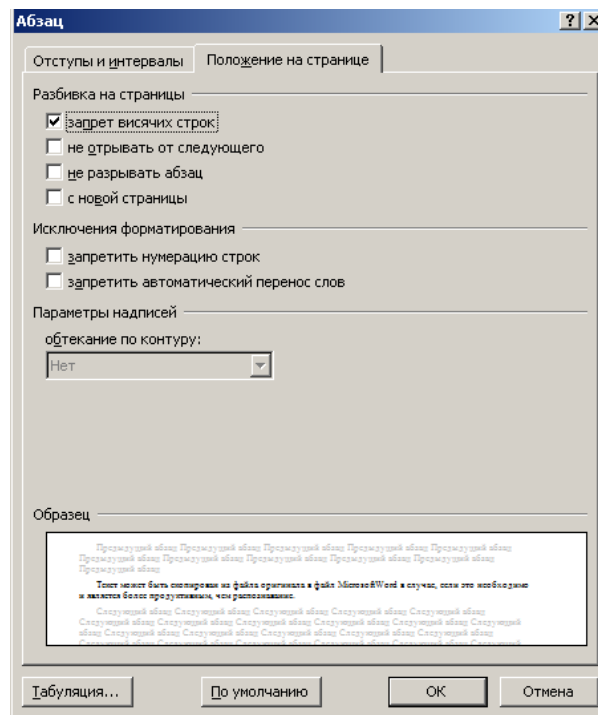
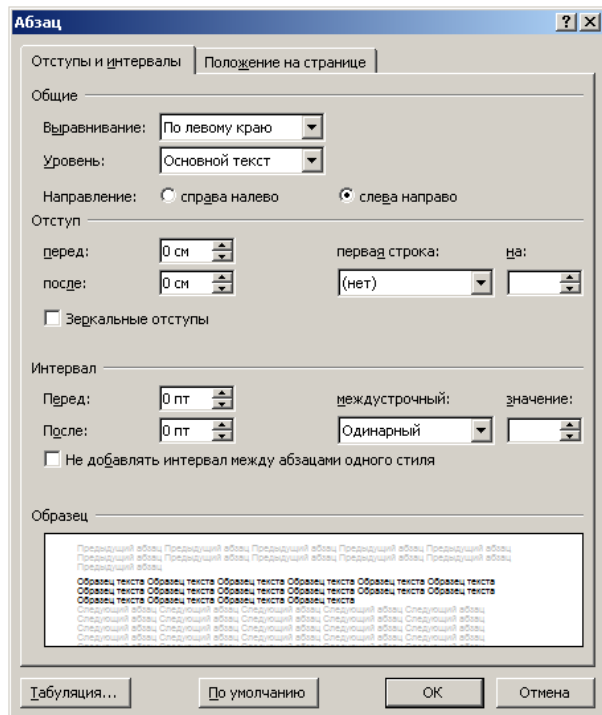
6. При применении иных программ распознавания следует выбрать настройки и выполнить действия, аналогичные приведенным в п.5 по результатам применения. Это замечание следует применять для всех дальнейших настроек/действий, если они указаны для ПО ABBYY FineReader.

2.1.4 Требования к тексту, полученному в результате набора или распознавания

1. Предварительная обработка текста, полученного в результате распознавания

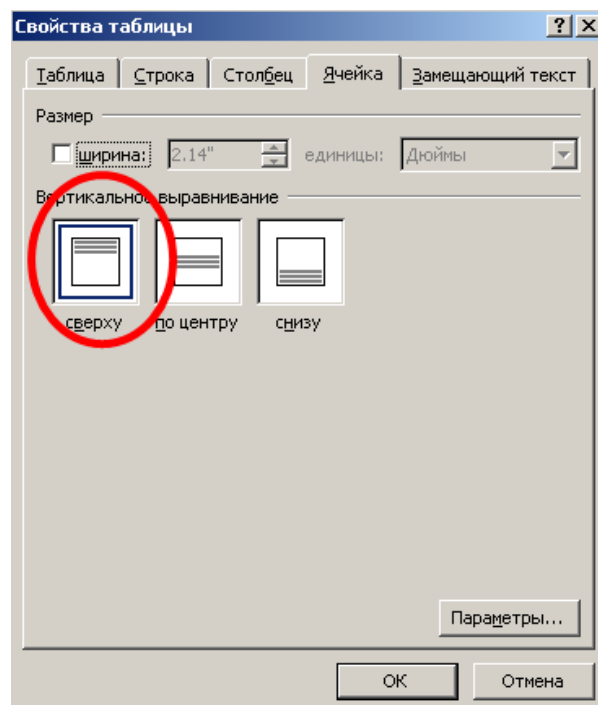
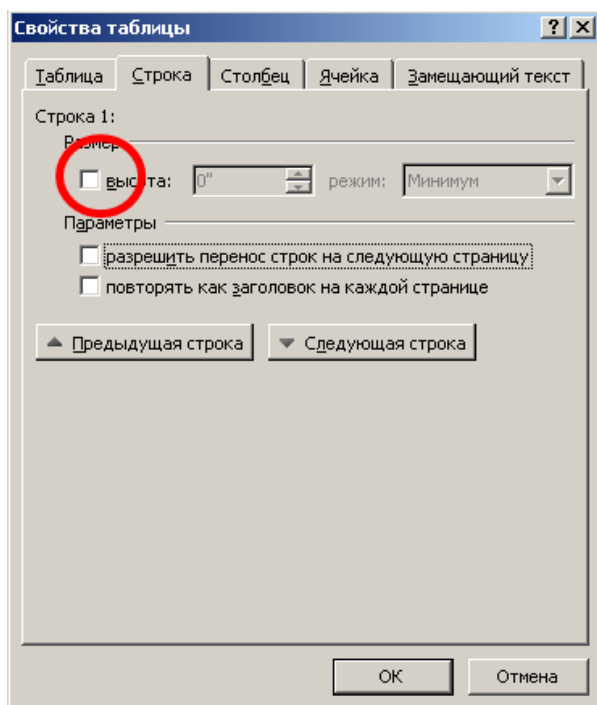
После распознавания текста при помощи программы ABBYY FineReader, проверки орфографии распознанного текста средствами этой программы (при необходимости) и сохранения его в файл формата DOCX (версия Microsoft Office не ниже 2010, какие-либо другие программы не применимы) следует проверить и установить:

1.1 Для всего текста следует установить следующие параметры форматирования абзацев:



1.2 Для всех таблиц в тексте следует установить следующие параметры форматирования:

(Показано для Microsoft Word 2010):



В свойствах таблицы должно быть отключено свойство «автоподбор размеров по содержимому», «обтекание» таблицы следует отключить.

При дальнейшей обработке текста, полученного в результате распознавания, программой Microsoft Word, допускается (рекомендуется) включить в Word режим автоматической проверки орфографии, что позволит выполнить дополнительную проверку орфографии, не выполняя при этом дополнительной отдельной операции.

Кроме указанных выше требований, текст должен удовлетворять приводимым в следующем пункте общим требованиям.

2. Общие требования

Текст, полученный в результате набора или распознавания, должен удовлетворять следующим требованиям к оформлению:

2.1 Электронный формат файла — Microsoft Word, формат документа — DOCX, если иное не указано.

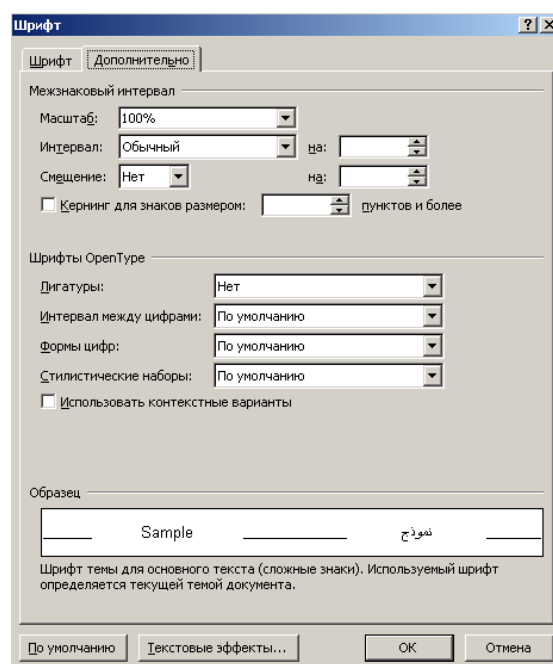
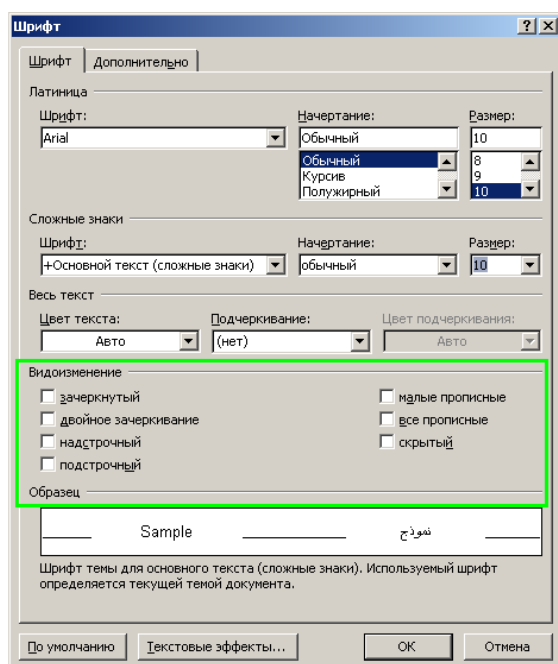
2.2 Размер бумаги в документе — A4 (A3 при необходимости таблиц с большим количеством данных), основное расположение листа — вертикальное. Поля документа — симметричные, по 1 дюйму (или 2,5 см).

2.3 Разделение на разделы отсутствует за исключением необходимости разворота листа, либо добавления листов другого размера бумаги (например, лист A3 горизонтальной ориентации при необходимости для таблицы в документе с основным размером бумаги «A4, вертикальное расположение»).

2.4 Применяемые шрифты — только Arial, Arial Narrow, Times New Roman или Courier New с соблюдением разделения шрифта на «рубленный»/«с засечками», если в оригинальном документе присутствует какое-либо выделение текста методом изменения типа шрифта. Шрифт Arial Unicode можно применять только в случае языков, содержащих символы, отсутствующие в обычных шрифтах, например для иероглифов.

2.5 Размер шрифта — 10-12 pt, единый для всего текста, кроме таблиц, в которых текст данного размера не может быть размещен без расстановки переносов. Масштабирование и изменение межбуквенного интервала не допускаются.

2.6 Для стиля «Обычный» следует указать шрифт того же размера и начертания, что приняты для основного текста; в качестве начертания шрифта следует выбирать только обычное начертание, не допускается выбор полужирного, курсива и т. п. В параметрах шрифта следует отключить какое-либо «оформление» («Малые прописные», «Все прописные», «Масштабирование» и т. п.), как показано на следующих иллюстрациях:



2.7 Весь текст должен быть расположен в одну колонку без оформления отступов абзацев. Расположение текста «как в оригинале» не требуется независимо от оформления исходного документа, если только иное не указано в задании. Следует отделять пустой строкой друг от друга текст, имеющий «разное значение», т. е. текст абзацев от текста надписей для иллюстраций и т. п. Слияние разных по смыслу текстовых единиц в одну не допускается. Деление текста на страницы, если это имеет смысл, следует выполнять символом конца страницы (Ctrl+Enter). Какие-либо иллюстрации (рисунки, блок-схемы и т. п.) не требуются, но текстовые надписи должны быть распознаны.

2.8 Межстрочный интервал должен быть равен единице. Допускается использовать единый для всего документа отступ после или до абзаца.

2.9 Все особые позиции табуляторов в тексте должны быть удалены.

2.10 В качестве параметров абзаца следует установить «Запрет висячих строк».

2.11 В качестве стиля выравнивания не допускается использовать выравнивание «по ширине», только «по левому краю влево» (основной стиль выравнивания, рекомендуется не пользоваться выравниванием «по правому краю», «по центру»).

2.12 Не рекомендуется использовать любые стандартные стили за исключением стиля «Обычный» и заголовочных стилей (только при необходимости автоматической сборки оглавления) [Заголовок N]. Не допускается создавать в документе стили. Стили, автоматически созданные программами распознавания, должны быть удалены из документа с тем, чтобы число стилей было минимально возможным. В случае, если программа распознавания создает сторонние стили, их следует удалить и назначить тексту стиль «Обычный».

2.13 Не следует выполнять любое форматирование при помощи пробелов и табуляторов:

- В тексте не должно быть подряд более одного пробела или табулятора
- Не должно встречаться сочетание «пробел»«табулятор» или «табулятор»«пробел».
- В середине предложения не должен встречаться знак «табулятор».

2.14 Следует отключить автонумерацию в любом виде (Маркированные списки, Нумерованные списки, и т. п.).

2.15 Символы «градус», «плюс-минус» и т. п. следует вставлять в текст при необходимости средствами вставки спецсимволов (Вставка → Символ) или вводом Alt-кода символа. Не допускается ввод символа градуса при помощи символов «0» (ноль), «O» (буква) и т. д., символа «±» при помощи подчеркивания знака «+» и т. п.

2.16 Если в тексте оригинала имеются буквы с диакритическими знаками (например, «Á», «Ê», «ù», «Ô» и т. п.) следует удостовериться, что они есть в распознанном тексте с соблюдением единообразия шрифта.

2.17 Если текст оригинала выделяется *курсивом*, *подчеркиванием*, *полужирным начертанием* или иным способом (цвет, надстрочный символ, и т. п.), то эти выделения должны быть выполнены. Если весь текст на странице выделен одинаковым цветом, то следует применять черный цвет текста (Авто) Цветные заливки (плашки) под текстом не нужны, если только не применяются для выделения ячеек в сложных таблицах или не выделяют отдельные слова в словосочетаниях/предложениях.

2.18 Не следует делить предложение с использованием знака конца параграфа (Enter) или конца строки (Shift+Enter), даже если в оригинале на этом месте начинается новая страница. Конец текста должен быть логически допустимым (конец предложения, конец абзаца). Если оригинал обрывается при отсутствии логического завершения (например, отсутствуют страницы с концом предложения, текст оригинала неразборчив из-за плохого качества и т. п.), следует уведомить об этом менеджера проекта (в электронном письме), а в текст внести фразу «обрыв оригинала» либо «неразборчиво», выделив ее при помощи инструмента «Цвет выделения текста».

В случае сомнений, являются ли несколько строк одной смысловой единицей (предложением), следует разделить их на строки при помощи символов Shift+Enter, в других случаях Shift+Enter не следует применять.

2.19 Запрещается расставлять переносы в тексте при помощи средств автоматической расстановки переносов, расстановка переносов вручную (соответствующим символом) не рекомендуется и допустима только в случаях крайней необходимости.

2.20 Элементы верхнего и нижнего колонтитулов набираются в основном тексте один раз в месте первого их появления, но не в колонтитулах.

2.21 Орфография набираемого текста должна быть проверена при помощи штатных средств проверки орфографии с учетом языка текста оригинала. Если в оригинале встречаются откровенные ошибки, их следует исправить, в случае каких-либо сомнений в написании слов рекомендуется оставить написание «как в оригинале», рекомендуется выделить такое слово при помощи инструмента «Цвет выделения текста» (в случае, если таких несловарных слов очень много (например, текст с медицинскими, химическими и т. п. терминами, которые отсутствуют в словарях проверки правописания), то можно их не выделять. Перед проверкой следует удостовериться в том, что выбранный в стиле «Обычный» и в качестве параметра для текста язык соответствует языку оригинала.

2.22 Все таблицы следует выполнять с помощью средств создания таблиц (Вставка → Таблица), даже если структура таблицы не обозначена явно линиями. Не следует выполнять разбиение текста по строкам таблицы при помощи «пустых строк» (клавиши Enter). Если текст логически делится на строки, то он должен быть размещен в отдельных строках таблицы. «Строка таблицы» не обязательно имеет видимые границы, например:

Строка 1	<i>Строка 1</i>
Строка 2	<i>Строка 2</i>
Строка 3	<i>Строка 3</i>
Строка 4	<i>Строка 4</i>

↓

Таблица!	
Строка 1	<i>Строка 1</i>
Строка 2	<i>Строка 2</i>
Строка 3	<i>Строка 3</i>
Строка 4	<i>Строка 4</i>

Не так!!!	
Строка 1¶	<i>Строка 1¶</i>
Строка 2¶	<i>Строка 2¶</i>
Строка 3¶	<i>Строка 3¶</i>
Строка 4¶	<i>Строка 4¶</i>

В данном примере символом «¶» обозначены символы, соответствующие нажатию на клавишу Enter. Пунктирными линиями в левой таблице обозначено разбиение на строки таблицы при невидимых границах.

Таблицы должны быть выполнены с сохранением структуры таблицы. В случае повышенной сложности структуры таблицы и при многократных объединенных по горизонтали ячеек допускается несоблюдение структуры таблицы, а только создание «базовой» структуры, с тем, чтобы окончательное форматирование было выполнено на этапе верстки (согласовать с менеджером), например:

Оригинальная структура таблицы:

xxx		xxx	xxx		xxx
xxx		xxx	xxx	xxx	Xxx
xxx	xxx	xxx	xxx	xxx	xxx
xxx		xxx		xxx	
xxx			xxx	xxx	xxx
xxx				xxx	
xxx	xxx	xxx		xxx	xxx
xxx	xxx	xxx		xxx	Xxx

Допускается (*только при предварительном согласовании*):

xxx			xxx	xxx			xxx
xxx		xxx	xxx		xxx	xxx	
xxx	xxx	xxx	xxx	xxx			xxx
xxx		xxx		xxx	xxx	xxx	
xxx					xxx		
xxx	xxx		xxx		xxx		xxx
xxx	xxx	xxx			xxx	xxx	

2.23 Следует сохранять соответствие «прописная-строчная» буква.

2.24 Запрещается набирать слова «вразбивку» т. е. с пробелами внутри слова, даже если подобное имеет место в оригинале, например: «ЗАГОЛОВОК», но не «З А Г О Л О В О К».

2.25 При использовании программ распознавания текста из файлов PDF возможно появление в распознанном тексте гиперссылок, которые невозможно устранить только настройками параметров программы распознавания. Следует обязательно удалять возможные гиперссылки (если иное не указано специально) в распознанных файлах формата Microsoft Word. Для этого в Microsoft Word надо выделить весь текст и нажать <Ctrl>+<Shift>+<F9>, после чего сохранить сделанные изменения.

3. Дополнительные требования к подготовке текста в презентациях формата Microsoft PowerPoint

При подготовке к переводу (распознаванию) файлов формата Microsoft PowerPoint следует дополнительно соблюдать следующие требования:

3.1 Не допускается изменять электронный формат исходной презентации с сторону «понижения версии» (конвертировать из формата *PPTX* в формат *PPT*), а также выполнять работу в несоответствующей версии программы, то есть работать с файлами в версии программы ниже PowerPoint 2010.

3.2 Следует применять только шрифты, которые имеют буквы кириллицы и стандартно присутствуют в ОС Microsoft Windows и ПО Microsoft Office (но не из коллекции шрифтов ПО Microsoft Office). Следует максимально возможно использовать шрифты Arial, Arial Narrow, Arial Black, Times New Roman или Courier New, Tahoma, Verdana, Calibri, Cambria с соблюдением разделения шрифта на «рубленный»/«с засечками», если в оригинальном документе присутствует какое-либо выделение текста методом изменения типа шрифта. Следует избегать использования не целых величин размера шрифта.

3.3 Таблицы, выполненные иллюстрациями, табуляторами, пробелами, отдельными текстовыми блоками, OLE-объектами для строк/колонок и т. п., следует выполнить в виде «таблицы» средствами PowerPoint (Вставка > Таблица).

3.4 Нераз редактируемые или имеющие недопустимое смысловое разделение (см. п. 2.18) надписи на иллюстрациях следует распознать или извлечь иным способом в соответствии с общими правилами и разместить при помощи инструмента «Надпись» на том же месте иллюстрации, с сохранением ориентации, тем же размером и начертанием (рубленный/с засечками/моноширинный) шрифта. Все «надписи» должны быть вне «полотна рисунка». С самой иллюстрации текст должен быть удален, допускается оставить оригинальный текст, сделав его невидимым, если любая попытка удаления приводит к неисправимым нарушениям изображения. Редактирование иллюстрации может быть выполнено любым способом с соблюдением условий:

- пропадание любых элементов иллюстрации, кроме собственно текста, недопустимо;
- цвета иллюстрации не должны изменяться;
- качество изображения не должно ухудшиться;
- размер файла PPTX не должен значительно увеличиться (в разы).

3.5 Следует сохранить все эффекты анимации, присутствующие в оригинальной презентации.

2.1.5 Требования к верстке в программе Microsoft Word

1. Оформление файла должно повторять оформление оригинала. При этом следует использовать только шрифты, имеющиеся в операционной системе Microsoft Windows «по умолчанию» (Arial, Arial Narrow, Arial Black, Times New Roman, Courier New, Tahoma, Verdana, Calibri, Cambria) с соблюдением разделения шрифта на «рубленный»/«с засечками», если в оригинальном документе присутствует какое-либо выделение текста методом изменения типа шрифта. Если специально не указано в задании, допускается изменение числа страниц для нормального размещения текста, при этом следует обратить внимание, что в случае присутствия в верстаемом тексте ссылок/указаний на номера страниц изменение расположения текста по страницам не допустимо. Следует избегать использования не целых величин размера шрифта.

2. Весь текст должен быть доступен для редактирования средствами программы Microsoft Word, при этом текст не должен содержаться в иллюстрациях и иных OLE-объектах, за исключением уравнений и элементов WordArt при, например, размещении текста по окружности. Следует учитывать, что после верстки объем текста в файле может быть изменен (может произойти как увеличение, так и уменьшение объема, например, при внесении редакторских правок в перевод), при этом не должно грубо нарушиться оформление документа.

- Любые таблицы следует выполнять со структурой «таблицы», а не при помощи табуляторов, независимо от того, ограничены ячейки таблицы какими-либо видимыми линиями или нет.

- Многоколоночный текст должен быть создан со структурой «колонок», а не, например, таблицы.

- Не следует размещать текст в требуемом месте при помощи пробелов.

- По возможности следует не применять более одного символа табулятора подряд.

- По возможности следует не применять нескольких символов конца абзаца (пустых строк).

- Следует не допускать окончания строки текста на символ табулятора или пробел, кроме случаев, когда этот символ выполняет функцию форматирования, например, табулятор с подчеркиванием до конца строки в бланках.

- Следует не допускать, по возможности, сочетания символа табулятора и пробела.

- Любой абзац должен являться целостной структурой, а не представлять собой набор отдельных строк, оканчивающихся символом конца абзаца (или конца строки), а также не должен принудительно прерываться при переходе на другую страницу, при необходимости разместить иллюстрацию в тексте абзаца и т. п., тем более подобным образом не следует разрывать предложение.

- По возможности следует избегать размещения текста по правому краю или по центру при помощи табулятора, следует использовать «выравнивание абзаца».

- Заголовки любого вида должны быть «привязаны» к абзацам, таблицам и т. п, следующим за ними (свойство абзаца «Не отрывать от следующего»). Это также относится к «строкам-заголовкам» таблиц. Аналогично, подписи, следующие за объектом (например, подпись к иллюстрации), должны быть «привязаны» к данному объекту.

- Не должны допускаться «висячие строки».

- Ячейки таблицы не должны разрываться (по возможности) при переходе со страницы на страницу. Если таблица располагается на нескольких страницах и имеет повторяющуюся строку-заголовок, то следует использовать специальную функцию, а не копировать строку несколько раз.

- Все текстовые блоки, содержащие надписи к какой-либо иллюстрации, графику и т. п. следует сгруппировать (если программа это позволяет).


– Любые графические элементы, создаваемые при помощи инструментов «Вставка» должны иметь логичную привязку к тексту. По возможности все графические элементы должны располагаться «в тексте», а не «поверх текста» (режим «обтекания» должен быть выключен) и не должны допускать «затекания» текста под элемент в процессе внесения правок.

– Не следует применять автоматическую расстановку переносов при помощи встроенных или внешних модулей. При необходимости расстановки переносов следует расставлять их вручную при помощи специальных символов переноса, а не символом дефиса.

– Символы, обозначающие величины единиц измерения, должны быть отделены от чисел «неразрывным» пробелом. Аналогично, символ неразрывного пробела должен стоять перед знаком тире, в сокращениях типа «т. д.», «т. п.», «в т. ч.» и подобных.

– Числовые величины не должны допускать разрывов числа: при разделении тысяч (если применяется) следует применять неразрывный пробел, при наличии в числе дефиса, например «2-1/2», следует применять неразрывный дефис.

– Символы следует вставлять при помощи операции Вставка -> Символ.

– Иллюстрация, являющиеся частью предложения (например: «...символ  означает...») должен быть вставлен, как часть предложения, а не как отдельная иллюстрация.

– Оглавление должно, по возможности, создаваться при помощи средств создания оглавления, а не вручную.

3. Элементы, повторяющиеся на нескольких страницах, следует делать при помощи верхних и нижних колонтитулов.

4. Неразделяемые или имеющие недопустимое смысловое разделение на отдельные строки надписи на иллюстрациях следует распознать или извлечь иным способом в соответствии с общими правилами и разместить в виде элемента «Надпись» на том же месте иллюстрации, с сохранением ориентации, тем же размером и начертанием (рубленный/с засечками/моноширинный) шрифта. Все «надписи» должны быть вне «полотна рисунка». С самой иллюстрации текст должен быть удален, допускается оставить оригинальный текст, сделав его невидимым, если любая попытка удаления приводит к неисправимым нарушениям изображения. Редактирование иллюстрации может быть выполнено любым способом с соблюдением условий:

- пропадание любых элементов иллюстрации, кроме собственно текста, недопустимо;
- цвета в иллюстрации не должны изменяться;
- качество изображения не должно ухудшиться;
- размер файла DOCX не должен значительно увеличиться (в разы).

5. Применяемое ПО: Microsoft Word, версии не ниже 2010.

2.1.6 Требования к названиям файлов

Название файла должно быть таким же, как название оригинального файла, либо, при разбиении файла на части) состоять из названия оригинального файла с дописанным к нему через дефис порядковым номером части, если иное не указано в задании.

_____ Конец рабочей инструкции