# Список сокращений

Это имитация исходного списка сокращений, составленного вручную при написании документа.

|  |  |
| --- | --- |
| **Аббревиатура** | **Расшифровка** |
| ABC | Шкала оценки Активации Внезапной Суперсилы |
| FDA | Food and drug administration / управление за качеством пищевых продуктов сша – неполная расшифровка, нестандартный стиль |
| Cintro | Какой-то новый параметр или новое написание аббревиатуры, которого еще нет в словаре |
| АД | Артериальное давление (нет в тексте) |
| МИО | Не должно быть добавлено в итоговую таблицу сокращений, так как не является аббревиатурой (в тексте в кавычках) |
| XIX | Не должно быть добавлено в итоговую таблицу сокращений, так как не является аббревиатурой (римские цифры) |

# Общие принципы работы алгоритма

Скрипт автоматически находит аббревиатуры в тексте и формирует список сокращений, но имеет ограничения (см. ниже), а также с тем, что результат во многом зависит от взаимодействия с пользователем. Аббревиатуры, которые могут попасть в итоговую таблицу списка сокращений (если пользователь не отказался от них) выделены в тексте зеленым, а те, что не должны – желтым. При каждом запуске алгоритм дополняет словарь введенными вручную новыми расшифровками. Основные этапы работы алгоритма:

1. Сканирование текста для поиска заглавных букв в слове
   * Поддерживаются как латиница (A-Z), так и кириллица (А-Я).
   * Минимум две заглавные буквы (например, ApoE, FDA).
   * Допустимы сочетания с цифрами или дефисами (например, ICAM-1, GlyT1).
2. Фильтрация
   * Игнорируются римские цифры и римские цифры с буквами (XIX, VIа, где **а** написано кириллицей, IIB и IIIa, где **а** – латинская буква, IV, IVb, VII).
   * Исключаются термины из стоп-листа (например, ПРОТОКОЛ КЛИНИЧЕСКОГО ИССЛЕДОВАНИЯ – написано заглавными буквами, но не является аббревиатурой).
   * Исключаются термины длинной более 8 букв, состоящие только из букв.
   * Слова в кавычках не рассматриваются как аббревиатуры («МИО», «RD-112»).
3. Остановка на разделе «Список литературы»
   * Если встречается заголовок (стиль Heading (любой) или выделенный жирным) с текстом «Список литературы» (включая вариации), скрипт прорускает этот раздел.

# Смешанные кириллические/латинские аббревиатуры

В документации могут встречаться неправильно написанные аббревиатуры: ATХ (первые две буквы – латиницей, последняя – кириллицей), СТСАЕ (все буквы кириллицей), ВГВ (последняя В – кириллицей, хотя вирусные гепатиты обозначаются латинскими буквами A, B, C, D, E и т.д.). Пользователю будет показана соответствующая верно написанная аббревиатура, **только** **если она есть в словаре**, а если нет (например, АTM c А – кириллицей, а остальные буквы - латиницей) – будет предложено ввести расшифровку, при этом, неверно написаная аббревиатура попадет в словарь. Таким образом, будьте внимательны – используйте правильный язык при написании сокращенной формы слова.

# Многозначные аббревиатуры

Аббревиатуры могут иметь несколько расшифровок (например, RR может означать Relative Risk, Risk Ratio или R-R Интервал). Программа покажет контекст и предложит выбрать подходящий вариант или добавить свой.

# Добавление новых аббревиатур и расшифровок

Когда в документе встречается новая аббревиатура, отсутствующая в словаре, программа предложит пользователю ввести расшифровку (описание). Процесс:

1. Программа покажет контекст в котором аббревиатура найдена (все куски текста.
2. Пользователь вводит расшифровку вручную. Если ничего не введено (нажата клавиша Enter), аббревиатура будет пропущена.
3. Программа проверяет, есть ли в словаре **похожие** описания.
   * Если найдены, программа спросит, действительно ли необходимо добавить новую аббревиатуру, так как ранее для той же цели использовалась иная.
   * Пользователь может выбрать одну из имеющихся пар или все же добавить новую аббревиатуру с введенной расшифровкой.

**Пример:**

Документ содержит новую аббревиатуру ТТг. Пользователь вводит расшифровку «Тиреотропный гормон (гипофиз)». Программа находит похожую пару (сопоставляя расшифровки):

1. ТТГ – Тиреотропный гормон

Пользователь может выбрать 1, чтобы добавить в таблицу сокращений пару ТТГ – Тиреотропный гормон, или нажать n, чтобы добавить новую пару ТТг – Тиреотропный гормон (гипофиз).

**Важно:** Если выбрана существующая аббревиатура (ТТГ), пользователю рекомендуется проверить текст документа и вручную заменить ТТг на ТТГ. Если пользователь добавит новую аббревиатуру (ТТг) и обновит словарь – обе записи останутся. Это может привести к необходимости ручной проверки для единообразия сокращений в будущем.

# Однобуквенные аббревиатуры

Аббревиатуры, состоящие из одной буквы, не могут быть найдены в тексте, так как не соответствуют правилу поиска «как минимум две заглавные буквы». Однако программа дополнительно ищет все сокращения из словаря в тексте. Если сокращения, состоящие из одной буквы были добавлены в словарь (например, А — пик скорости фазы систолы предсердий, M — Males (самцы) или Mean (среднее арифметическое)), программа спросит пользователя, нужно ли добавить каждое из них в список сокращений.

# Сокращения в таблицах

|  |  |
| --- | --- |
| **Рост** | **ИМТ (кг/м2)** |
| 25 | 75,5 |
| 85 | 1,5 |
| **Примечание:**  ИМТ – индекс массы тела. | |

# Ограничения

1. Если сокращение ни разу не использовалось в тексте, а есть только в исходной таблице сокращений, в сгенерироавнной скриптом таблице оно не появится.
2. Алгоритм исключает ограниченный список терминов. Если в тексте окажутся другие слова, которые систематически пишутся заглавными буквами, программа продолжит распознавать их как сокращения и предлагать ввести расшифровку. Чтобы добавить такие слова в список исключений, необходимо уведомить разработчика.
3. Могут встречаться сложные (не удовлетворяющие условиям поиска) аббревиатуры, например, фармакокинетические параметры: Vd, Cmax, T1/2. Указанные ФК параметры уже есть в словаре и будут автоматически найдены программой. Однако новые сложные аббревиатуры, такие как Cinf или Tbc, не будут обнаружены.
4. Аббревиатуры из двух слов (например, NCI-CTC AE) программа распознает как отдельные сокращения. Если пользователь введет расшифровку для каждого компонента, а в словаре уже есть полная форма, в таблице окажутся все три записи. В настоящий момент в словаре присутствуют расшифровки для компонентов «NCI» и «CTCAE», но нет варианта написания «NCI-CTC» и нет аббревиатеры AE. Поэтому в таблице окажутся только расшифровки отдельно для NCI-CTC и AE (если если этот шаг не был пропущен нажатием Enter вместо ввода текста).
5. При подготовке документа следует учитывать, что на данный момент единственный способ обработки аббревиатур, состоящих из нескольких слов – обращение к разработчику с просьбой внести их в словарь.
6. Скрипт не проверяет, соответствует ли найденная аббревиатура тому значению, которое имел в виду автор. Если в словаре есть только одна возможная расшифровка, программа автоматически выберет её, даже если в контексте требуется иной вариант. Обращение к словарю для ознакомления с имеющимися расшифровками и к разработчику с просьбой обновить словарь заранее позволит избежать таких ситуаций.
7. Если одна и та же аббревиатура встречается в словаре и тексте в разных форматах, например, AUC0-24 и AUC(0-24), все варианты попадут в итоговую таблицу (auc в данном случае тоже попадет в таблицу, так как имеется в словаре и будет найден поиском всех аббревиатур словаря в тексте).
8. Алгоритм прекращает поиск при обнаружении заголовка «Список литературы» (Heading 1). Если раздел оформлен в другом стиле или под другим названием, скрипт продолжит поиск и может добавить в таблицу сокращений аббревиатуры из литературных ссылок.
9. Если при вводе расшифровки пользователь допустит ошибку, скрипт не сможет её исправить. В этом случае потребуется ручное редактирование и словаря и таблицы сокращений.

# Проверки

Если в документе уже есть вручную составленная таблица сокращений, программа проведёт сравнение и выявит различия между существующей таблицей и автоматически сгенерированной.

Результаты проверки включают:

* Аббревиатуры, которые присутствуют в исходной таблице, но не были найдены алгоритмом.
* Аббревиатуры, которые обнаружены программой, но отсутствуют в существующей таблице.
* Разные аббревиатуры с одинаковой расшифровкой в итоговой таблице.
* Разные расшифровки для одной и той же аббревиатуры (если скрипт покажет такие, обратитесь, пожалуйста, к разработчику – скорее всего в ошибка в скрипте или словаре).

# Список литературы

Аббревиатуры, использованные в списке литературы не добавляются в таблицу сокращений.

1. World Health Organization. WHO Guidelines on ...