Lingtrain Aligner

Sergei Averkiev

апр. 21, 2021

Быстрый старт

1 Подготовка документов **3**

2 Загрузка документов **5**
2.1 Выбор языка 5
2.2 Загрузка 5
2.3 Удаление документов 7

3 Вспомогательный перевод **9**
3.1 Варианты создания перевода 9
3.2 Загрузка перевода 11

4 Создание выравнивания **13**

5 Редактирование **15**
5.1 Процесс выравнивания 16
5.2 «Потерянные» строки 20

6 Модели **23**

Lingtrain Aligner — это веб-приложение, которое поможет вам выровнять два текста на разных языках. Сейчас поддерживается 10+ языков, для любой пары из которых доступно выравнивание.

Первоначальная обработка происходит автоматически при помощи *предобученных моделей машинного обучения*. Пользователь валидирует результат и дорабатывает полученный параллельный корпус до приемлемого качества.

Глава 1

Подготовка документов

Приложение принимает на вход «сырые» документы в txt формате. Для улучшения качества проследите, чтобы тексты начинались с одного и того же места (например, с одной и той же главы, если речь идет о художественном произведении и его переводе).

Для примера возьмем рассказ Максима Горького «Макар Чудра» на русском и китайском языках.

makar_ru.txt

```
С моря дул влажный, холодный ветер, разнося по степи
задумчивую мелодию плеска набегавшей на берег волны и
шелеста прибрежных кустов. Изредка его порывы приносили
с собой сморщенные, желтые листья и бросали их в костер,
раздувая пламя; окружавшая нас мгла осенней ночи
вздрагивала и, пугливо отодвигаясь, открывала на миг слева
- безграничную степь, справа - бесконечное море и прямо
против меня - фигуру Макара Чудры, старого цыгана, - он
сторожил коней своего табора, раскинутого шагах в
пятидесяти от нас.
Не обращая внимания на то, что холодные волны ветра,
распахнув чекмень, обнажили его волосатую грудь и
безжалостно бьют ее, он полулежал в красивой, сильной позе,
лицом ко мне, методически потягивал из своей громадной
трубки, выпускал изо рта и носа густые клубы дыма и,
неподвижно уставив глаза куда-то через мою голову в мертво
молчавшую темноту степи, разговаривал со мной, не умолкая
и не делая ни одного движения к защите от резких ударов
ветра.
...
```

makar_zh.txt

```
.
```

(continues on next page)

(продолжение с предыдущей страницы)

```
.
 50 ‘‘ ’’ ‘«’ ‘‘ ‘‘ ...
```

В данном примере тексты приведены «как есть», без какого-либо разбиения по предложениям.

Глава 2

Загрузка документов

2.1 Выбор языка

Перед загрузкой выберите языки из раскрывающегося списка на панели в верхней части приложения. Затем, щелкнув на скрепку, выберите нужный файл и нажмите кнопку Upload. После успешной загрузки название файла появится в верхней части секции. Убедитесь, что язык документа соответствует языку, выбранному на верхней панели. Повторите процедуру для второго документа.

2.2 Загрузка

Рекомендуется загружать текст на исходном языке в левую колонку, а текст на языке перевода – в правую, так как программа берет за основу предложения именно из левой колонки и подбирает к ним наиболее подходящие по смыслу из правой. К примеру, для выравнивания рассказа Максима Горького «Макар Чудра» и его перевода на китайский язык нужно загрузить текст на русском языке в левую колонку, а текст на китайском – в правую.

Russian Chinese

Preview

Documents are splitted by sentences using language specific rules.

makar_ru.txt

377 lines

1	С моря дул влажный, холодный ветер, разнося по степи задумчивую мелодию плеска набегавшей на берег волны и шелеста прибрежных кустов.
2	Изредка его порывы приносили с собой сморщенные, желтые листья и бросали их в костер, раздувая пламя; окружавшая нас мгла осенней ночи вздрагивала и, пугливо отодвигаясь, открывала на миг слева — безграничную степь, справа — бесконечное море и прямо против меня — фигуру Макара Чудры, старого цыгана, — он сторожил коней своего табора, раскинутого шагах в пятидесяти от нас.
3	Не обращая внимания на то, что холодные волны ветра, распахнув чекмень, обнажили его волосатую грудь и безжалостно бьют ее, он полулежал в красивой, сильной позе, лицом ко мне, методически потягивал из своей громадной трубки, выпускал изо рта и носа густые клубы дыма и, неподвижно уставив глаза куда—то через мою голову в мертво молчавшую темноту степи, разговаривал со мной, не умолкая и не делая ни одного движения к защите от резких ударов ветра.

makar_zh.txt

431 lines

1	马加尔·楚德拉一阵潮湿而寒冷的风从海上吹来，把拍打着堤岸的波浪发出的溅水声和沿岸灌木丛发出的飒飒声组成的低沉的旋律吹散在草原上。
2	劲风不时地卷起枯皱的黄叶，把它们投到篝火堆里，燃起了火焰。
3	我们周围的秋夜的黑幕颤动起来，慌忙地退开了，顿时，左面露出了辽阔的草原，右面是无垠的大海，而正面对着我的是马加尔·楚德拉这位老茨冈人的身影。
4	他在看守着他们游牧队的马群，他们的屯宿地离我们只有50步左右。
5	寒风撕开了他身上的捷克曼上衣，无情地扑打着他那毛茸茸的胸膛，但是他一点儿也不在乎。
6	他斜躺着，姿态很美，现在刚劲有力的样子，脸对着我，有节奏地吸着他的大烟斗，浓烟从他的鼻孔和口腔里冒了出来。
7	他的目光从我的头上越过，凝视着死一样沉寂的黑暗的草原，不停地和我交谈着，对冷风的无情袭击没有采取任何抵抗行动。
8	你就这样流浪吗？

В каждую из колонок можно загрузить несколько документов. Названия загруженных файлов будут появляться в верхней части секции, а выбранный документ – подсвечиваться серым. Для выбора нужного документа из списка загруженных достаточно просто щелкнуть по нему. При одновременной работе с несколькими документами перед созданием выравнивания убедитесь, что файлы выбраны правильно.

Russian Chinese

Documents

makar_ru.txt

detstvo_ru.txt

Upload

Upload raw Russian document in txt format.

detstvo_ru.txt

UPLOAD

makar_zh.txt

detstvo_zh.txt

Upload

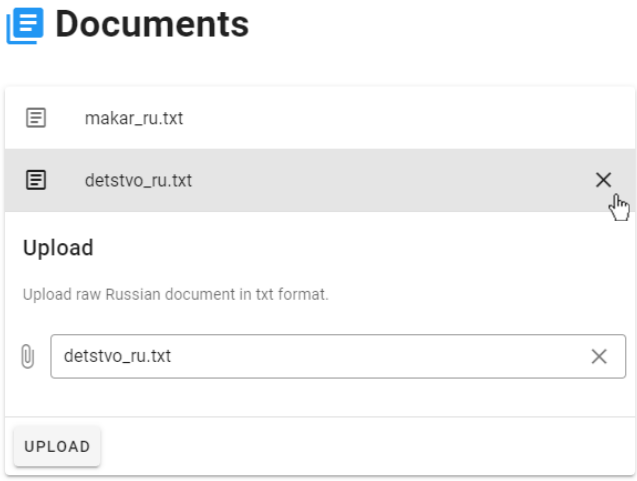
Upload raw Chinese document in txt format.

detstvo_zh.txt

UPLOAD

2.3 Удаление документов

Для того, чтобы удалить загруженный документ, нужно нажать на значок крестика, который появляется при наведении курсора на название файла.

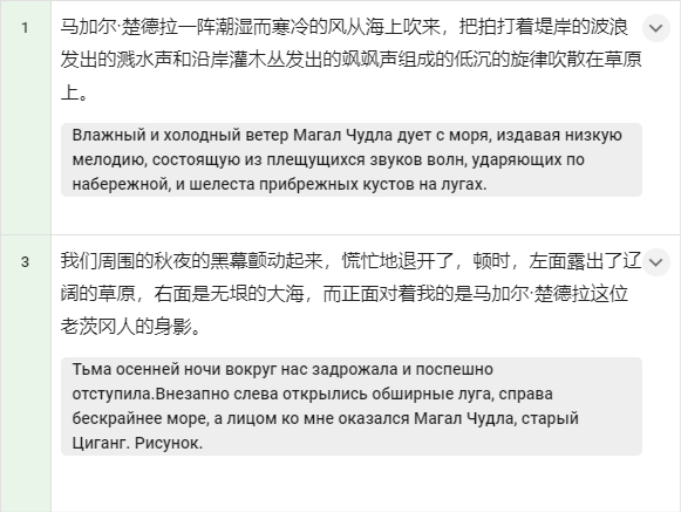


Обратите внимание: после загрузки документы закрепляются за выбранными на верхней панели языкам. Если после загрузки документов Вы обнаружили, что неправильно установили языки, удалите загруженные файлы, выберите языки правильно и загрузите документы заново.

Глава 3

Вспомогательный перевод

Для удобства работы по желанию можно добавить к тексту на иностранном языке вспомогательный машинный перевод, который будет отображаться под каждой строкой текста в области редактирования.



3.1 Варианты создания перевода

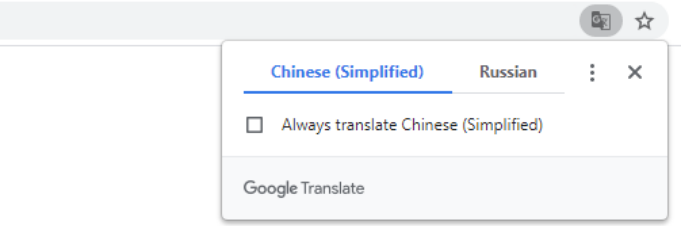
3.1.1 Первый вариант

Чтобы создать файл со вспомогательным переводом, перейдите к секции Preview и скачайте текст на иностранном языке с помощью кнопки Download. Обратите внимание: в отличие от исходного файла текст в скачанном документе разбит по строкам. Переведите текст с помощью любого из доступных Вам сервисов или программ машинного перевода (например, Google Переводчик) и сохраните в формате txt; убедитесь, что разбиение по строкам сохранилось.

3.1.2 Второй вариант

Также можно воспользоваться встроенным в браузер Chrome переводчиком сайтов. Для этого нажмите на кнопку Open, при этом текст откроется в браузере без скачивания (это важно так как Chrome не позволяет переводить локальные файлы).

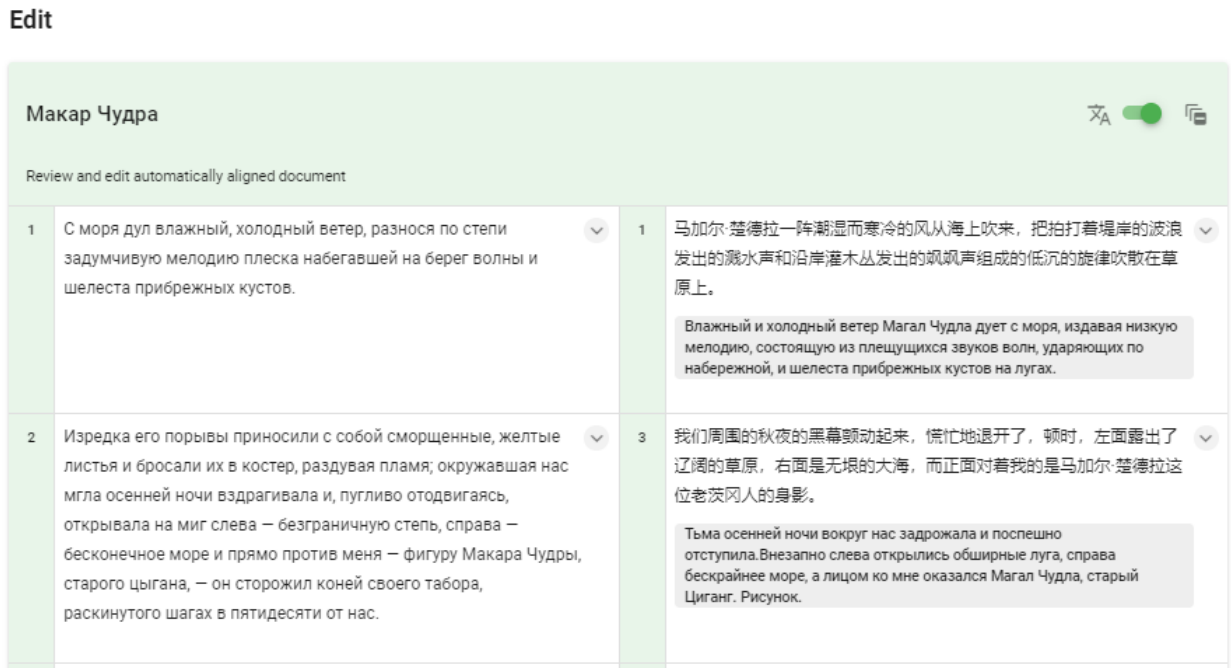
Chrome предложит перевести страницу на русский язык. Переведенный текст нужно сохранить к себе на компьютер и перейти к следующему шагу.



3.2 Загрузка перевода

Для загрузки вспомогательного перевода нажмите кнопку **Add translation hint** под текстом в секции Preview. Откроется область Upload translation; выберите файл с переводом, нажав на значок скрепки, после чего нажмите кнопку Upload.

Затем создайте выравнивание как обычно. После выравнивания первой части текста в области редактирования Edit необходимо передвинуть ползунок вправо. Вспомогательный перевод появится под строками иностранного текста. Отключить вспомогательный перевод можно в любой момент, передвинув ползунок влево.



Обратите внимание, что файл со вспомогательным переводом необходимо загрузить до начала выравнивания.

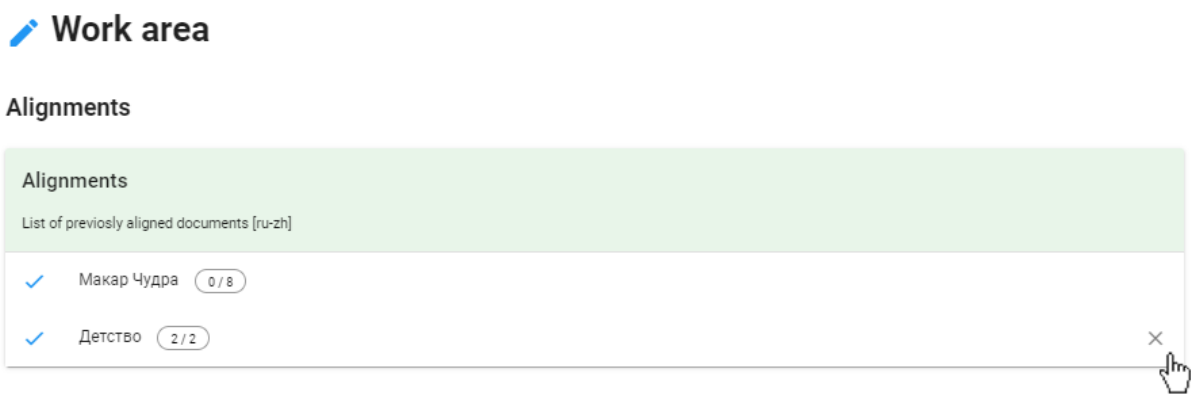
Глава 4

Создание выравнивания

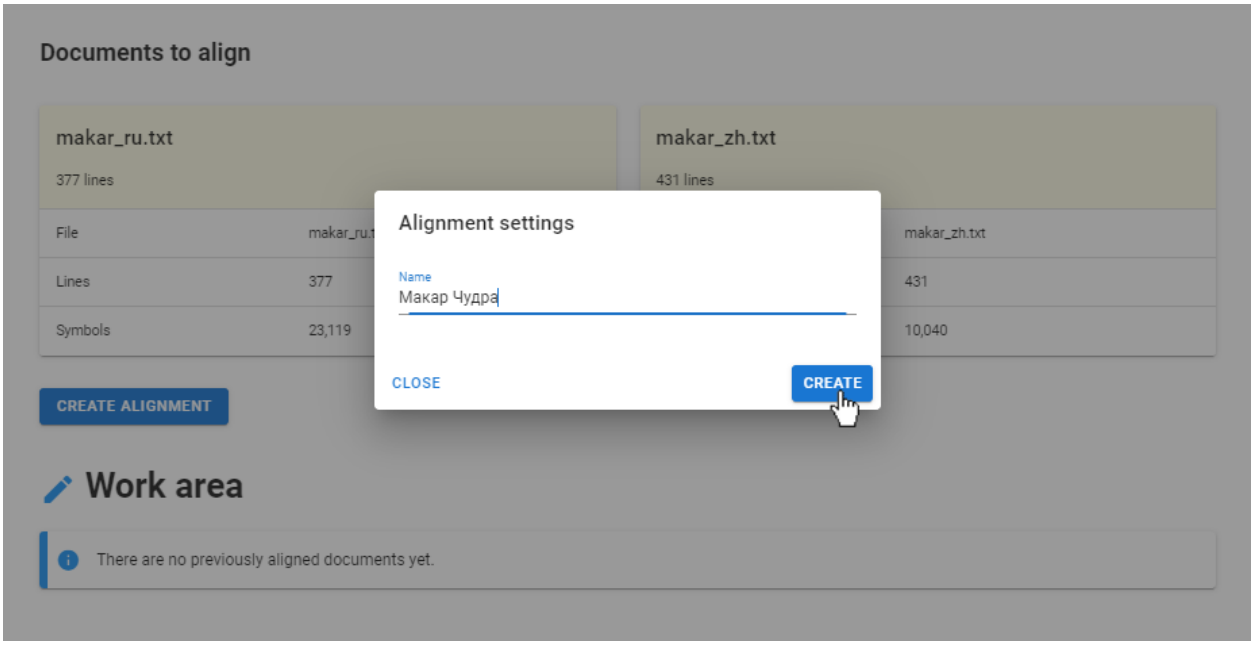
Для создания выравнивания выберите пару загруженных ранее документов на разных языках и нажмите на кнопку **Create alignment**. Появится диалоговое окно, в котором необходимо ввести название для выравнивания.

Нажмите кнопку **Create**. В секции Work area появится новое выравнивание. Программа автоматически разбивает тексты на части по 50 строк – батчи. Два числа через дробь справа от названия выравнивания показывают, на сколько батчей разбит текст и сколько из них уже выровнено. В нашем примере рассказ Максима Горького «Макар Чудра» разбит на 8 батчей, из них сейчас выровнено 0.

При работе с несколькими парами текстов для каждой пары можно создавать отдельное выравнивание. Чтобы выбрать нужное выравнивание, достаточно просто щелкнуть по нему. Чтобы удалить созданное выравнивание, нужно нажать на значок крестика, который появляется при наведении курсора.



Чтобы выровнять первую часть, нажмите кнопку **Align next batch**. Отредактируйте первый батч в секции *Edit* и нажмите кнопку Align next batch, чтобы начать работу со следующей частью. Переходите от батча к батчу, пока не отредактируете корпус целиком.



Глава 5

Редактирование

Непосредственная работа с текстами ведется после *создания выравнивания* в секции Edit. Область редактирования представляет собой таблицу с двумя колонками: в левой колонке по порядку расположены предложения текста на исходном языке, а в правой программа автоматически подбирает к ним наиболее подходящие кусочки текста на языке перевода. Числа слева от ячеек указывают на порядковый номер предложения в тексте. В нашем примере программа сопоставила первое русское предложение с первым китайским, а для второго предложения русского текста подобрала третье предложение китайского текста как наиболее подходящее.

Изображение 1

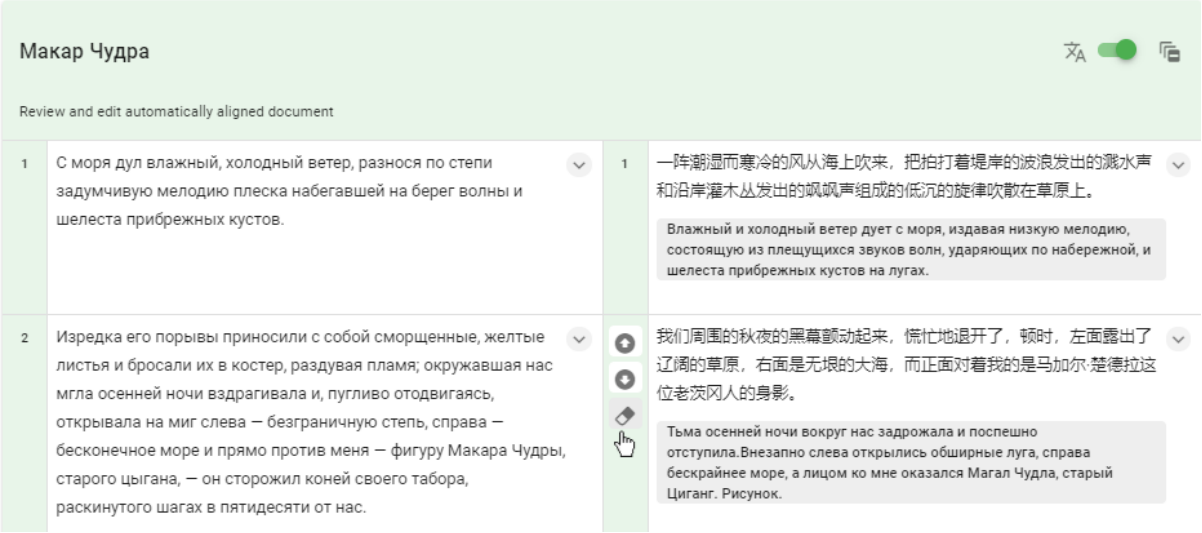
5.1 Процесс выравнивания

В ходе ручного выравнивания текстов необходимо проверить, правильно ли программа сопоставила пары предложений и, если нужно, внести изменения. Редактирование производится с помощью нескольких кнопок:

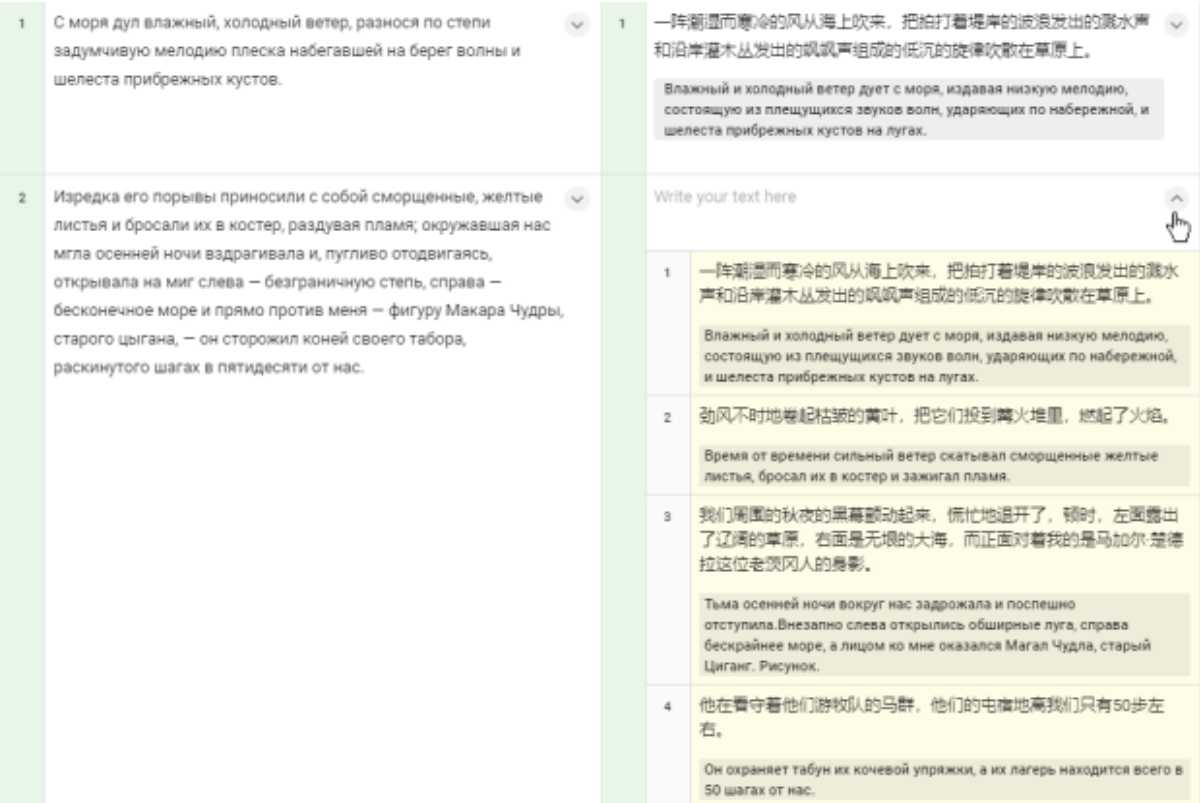
Кнопка	Описание
	скопировать в предыдущую строку/добавить к текущей строке
	скопировать в следующую строку
	удалить содержание ячейки
+ строка	добавить строку
удалить	удалить строку

Рассмотрим несколько примеров.

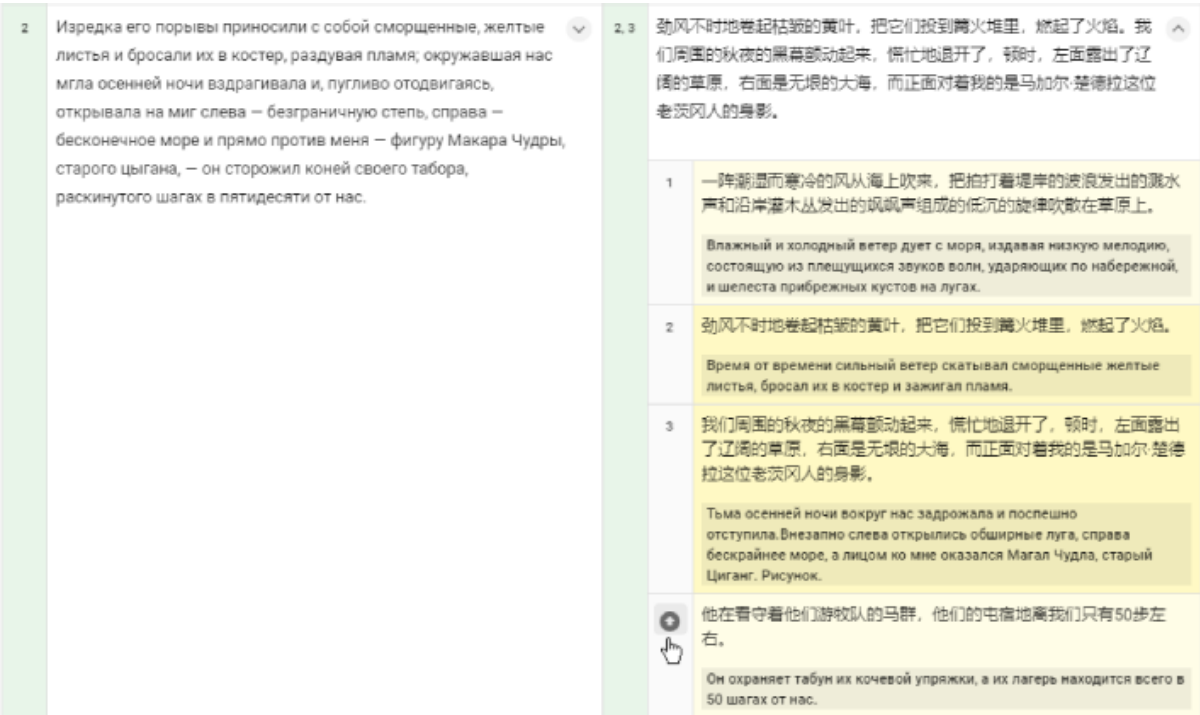
В этом примере первая пара предложений сопоставлена правильно, а вот второе русское предложение переведено на китайский не полностью – не хватает частей «Изредка его порывы приносили с собой сморщенные, желтые листья и бросали их в костер, раздувая пламя» и «он сторожил коней своего табора, раскинутого шагах в пятидесяти от нас» (*Изображение 1*). Чтобы добавить недостающие кусочки, для начала очистим всю ячейку целиком. Наведите курсор на номер предложения: появятся три значка – стрелка вверх, стрелка вниз и ластик. Нажмите на ластик, чтобы очистить ячейку.



Теперь нажмем на стрелку в правой части ячейки. Откроется список кандидатов – предложений,расположенных в тексте рядом с текущим.



Мы видим, что в ячейку нужно поместить предложения №2, №3 и №4. Чтобы поместить предложение в предыдущую ячейку, наведите курсор на номер предложения и нажмите на значок стрелки, указывающей вверх. Поместите предложения №2, №3 и№4 по порядку.



Закройте список кандидатов, нажав на стрелку в правой части ячейки; строка готова.

Макар Чудра

Review and edit automatically aligned document

1	С моря дул влажный, холодный ветер, разнося по степи задумчивую мелодию плеска набегавшей на берег волны и шелеста прибрежных кустов.	1	一阵潮湿而寒冷的风从海上吹来，把拍打着堤岸的波浪发出的溅水声和沿岸灌木丛发出的飒飒声组成的低沉的旋律吹散在草原上。 Влажный и холодный ветер дует с моря, издавая низкую мелодию, состоящую из плещущихся звуков волн, ударяющих по набережной, и шелеста прибрежных кустов на лугах.
2	Изредка его порывы приносили с собой сморщенные, желтые листья и бросали их в костер, раздувая пламя; окружавшая нас мгла осенней ночи вздрагивала и, пугливо отодвигаясь, открывала на миг слева — безграничную степь, справа — бесконечное море и прямо против меня — фигуру Макара Чудры, старого цыгана, — он сторожил коней своего табора, раскинутого шагах в пятидесяти от нас.	2, 3, 4	劲风不时地卷起枯皱的黄叶，把它们投到篝火堆里，燃起了火焰。我们周围的秋夜的黑幕颤动起来，慌忙地退开了，顿时，左面露出了辽阔的草原，右面是无垠的大海，而正面对着我的是马加尔·楚德拉这位老茨冈人的身影。他在看守着他们游牧队的马群，他们的屯宿地离我们只有50步左右。

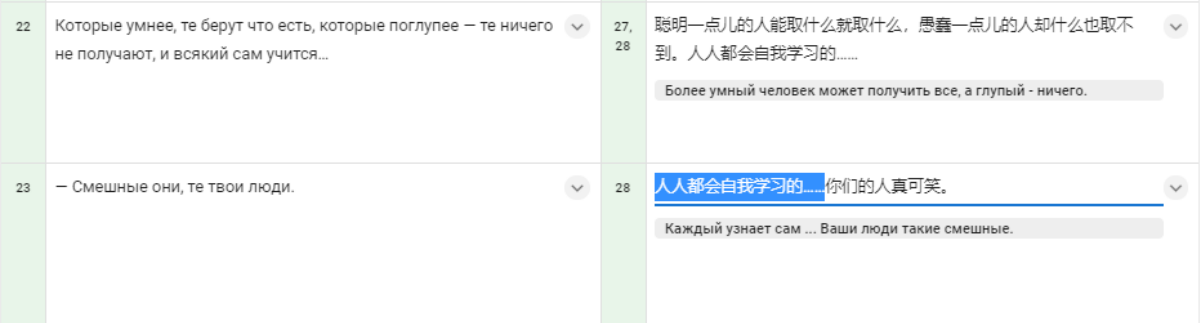
В следующем примере в строке №27 китайского текста не хватает части «и всякий сам учится...»: этот кусочек попал в строку №28.

22	Которые умнее, те берут что есть, которые поглупее — те ничего не получают, и всякий сам учится...	27	聪明一点儿的人能取什么就取什么，愚蠢一点儿的人却什么也取不到。 Более умный человек может получить все, а глупый - ничего.
23	— Смешные они, те твои люди.	28	人人都会自我学习的.....你们的人真可笑。 Каждый узнает сам ... Ваши люди такие смешные.

Удобнее всего будет добавить строку №28 к строке №27 и стереть лишнюю часть. Чтобы добавить строку к предыдущей, наведите курсор на номер строки и нажмите на стрелку вверх.

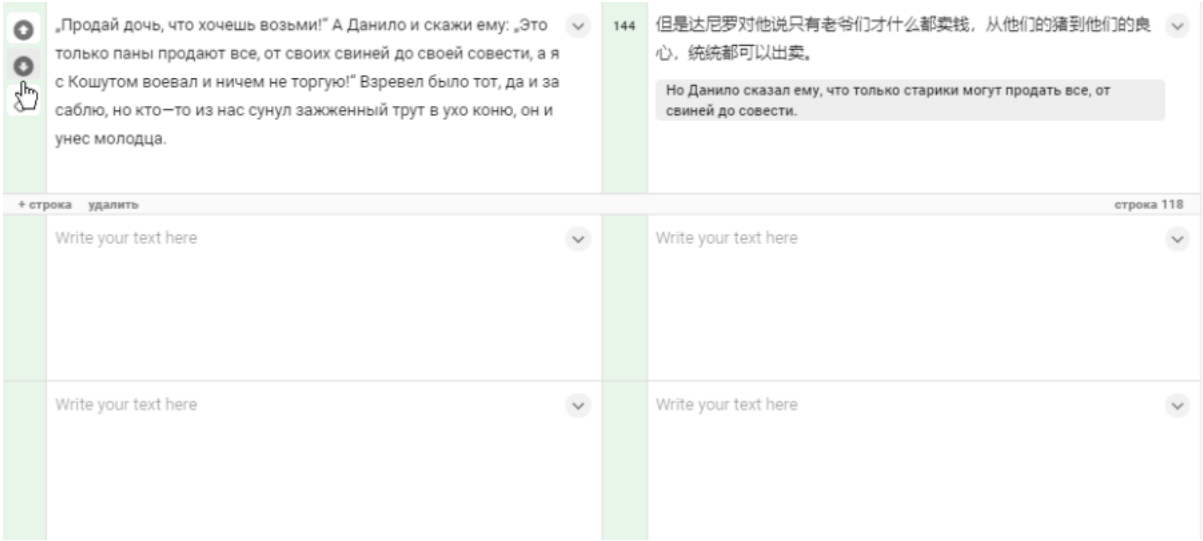
22	Которые умнее, те берут что есть, которые поглупее — те ничего не получают, и всякий сам учится...	27	聪明一点儿的人能取什么就取什么，愚蠢一点儿的人却什么也取不到。 Более умный человек может получить все, а глупый - ничего.
23	— Смешные они, те твои люди.		人人都会自我学习的.....你们的人真可笑。 Каждый узнает сам ... Ваши люди такие смешные.

Теперь осталось только стереть лишнее в обеих строках. Текст в ячейках можно стирать, допечатывать, копировать и вставлять, как в обычном текстовом редакторе.



В следующем примере в русскую строку №120 попало сразу несколько предложений, потому что в оригинальном документе не были проставлены точки. Строку нужно разбить на отдельные предложения. Для этого наведите курсор на нижнюю часть ячейки: появятся кнопки «+ **строка**» и «**удалить**», а также номер строки.

Добавим две строки и скопируем содержание строки №120 в каждую из них, нажав на стрелку вниз.

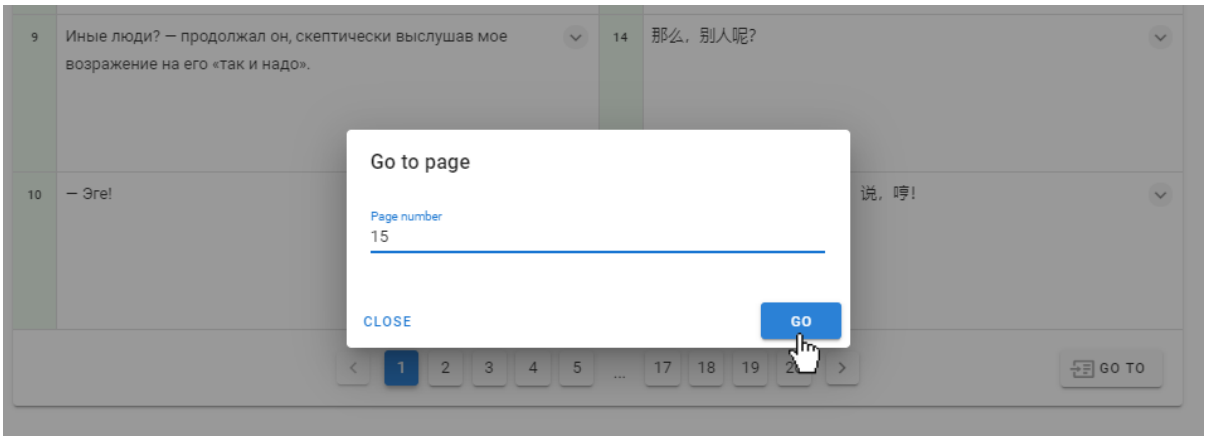


Затем удалим лишнее в каждой из ячеек и подберем к ним китайские соответствия.

120	„Продай дочь, что хочешь возьми!".	143	把你女儿卖给我，随便你要什么都可以！
120	А Данило и скажи ему: „Это только паны продают все, от своих свиней до своей совести, а я с Кошутом воевал и ничем не торгую!".	144, 145	但是达尼罗对他说只有老爷们才什么都卖钱，从他们的猪到他们的良心，统统都可以出卖。可是我这个人是跟噶苏特打过仗的，我不做什么买卖！
120	Взревел было тот, да и за саблю, но кто—то из нас сунул зажженный трут в ухо коню, он и унес молодца.	146, 147	大阔老大声吼叫着，伸手去抽他的佩刀。这时候，我们中间有人把点燃的火绒塞进了他坐骑的耳朵里，马跳起来，驮着他跑走了。

При необходимости строки также можно удалять, нажав на кнопку «**удалить**» в нижней части строки.

Таким образом переходите от строки к строке, пока не отредактируете весь текст. Программа показывает по 10 строк на одной странице. В нижней части области редактирования расположены номера страниц. Текущая страница подсвечивается синим; чтобы перейти на нужную страницу, просто нажмите на ее номер или воспользуйтесь стрелками. Кнопка **Go to** to позволяет быстро перейти к той или иной странице. К примеру, Вам нужно просмотреть строку №141 исходного текста; она находится на странице 15. Нажмите кнопку **Go to**, введите число 15 и нажмите **Go**.

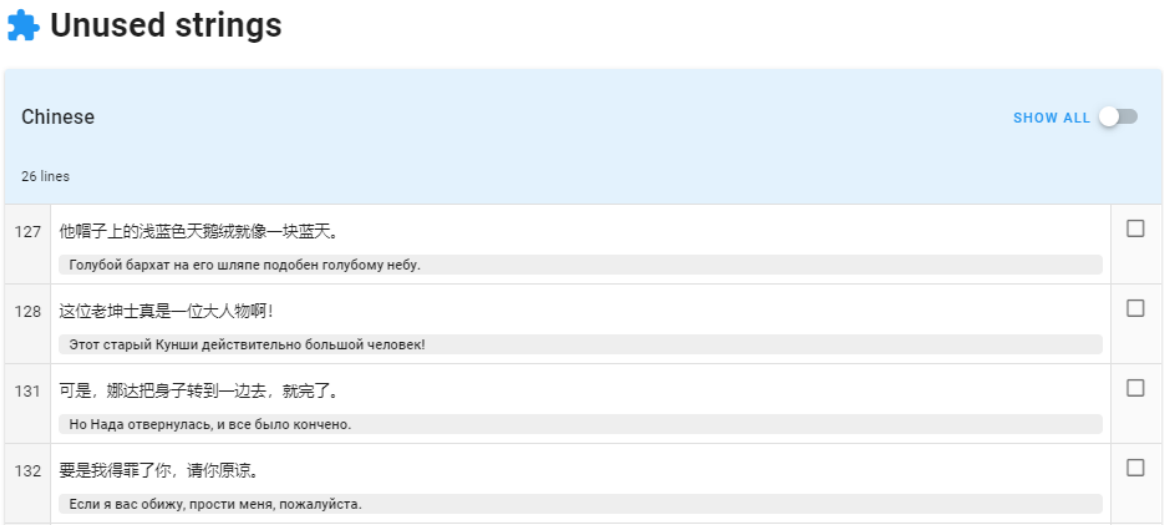


5.2 «Потерянные» строки

В процессе выравнивания искать несоответствия удобнее всего по номерам строк в левой колонке. Ищите пропущенные номера: например, здесь после строки №226 сразу идет строка №228 – очевидно, что строка №227 потерялась при автоматическом выравнивании.

193	Все он мог сделать с человеком, и все любили его, крепко любили, только Радда одна не смотрит на парня; и ладно, коли бы только это, а то еще и подсмеивается над ним.	226	他能够叫人做任何事情，大家都爱他，并且爱得很深，只有娜达一人连看都不看他一眼。 Он может просить людей сделать что угодно. Все его любят и сильно любят. Только Нада даже не смотрит на него.
194	Крепко она задела за сердце Зобара, то—то крепко!	228	她深深地刺痛了佐巴尔的心，啊，深深地！ Она пронзила сердце Зобала глубоко, ах, глубоко!

Для удобства работы все незадействованные строки попадают в секцию **Unused strings**, которая расположена сразу под областью редактирования. Строки собраны в таблицу по порядку размещения в тексте; в шапке таблицы указаны язык текста, в котором есть «потерянные» строки, а также их количество.



Когда в процессе ручного выравнивания строка оказывается задействована, она исчезает из таблицы. Ненужные строки, которые при ручном выравнивании оказались лишними, можно помечать с помощью галочки в правой колонке. В шапке таблицы также расположен ползунок **Show all**: когда он сдвинут влево, строки, отмеченные галочками, скрыты. Чтобы показать все незадействованные строки, достаточно передвинуть ползунок вправо. В нашем примере строки №12, №79 и №105 помечены как незадействованные и подсвечены серым; если передвинуть ползунок влево, эти строки отображаться не будут.

После завершения работы с документами рекомендуем всегда просматривать секцию Unused strings, чтобы убедиться, что никакие строки не потерялись в процессе выравнивания. Скачать готовые тексты можно в секции Corpora, нажав на кнопку Download: каждый текст по отдельности в формате txt или сразу оба в формате tmx.

Глава 6

Модели

Про модели машинного обучения.