SDL BeGlobal Trainer:

как очень быстро настроить МП, имея хорошую (-ие) ТМ

SDL BeGlobal Trainer: основные факты

- SDL BeGlobal Trainer - это SaaS-приложение (и компонент SDL BeGlobal)
- Пользователи могут применять его для настройки и/или создания языковых пар по определенным предметным областям
- В основе решения лежит технология, многие годы использовавшаяся в Language Weaver/SDL, и представленная теперь в рамках нового удобного пользовательского интерфейса
- Обучение основано на обработке параллельных данных (например, в формате TM); для достижения приемлемого результата потребуется наличие, как минимум, 200 000 исходных слов или 15 000 предложений
- В среднем, требуется от 24 до 36 часов для настройки, в зависимости от загруженности сервера, объема используемых данных и т.д.
- Кто может быть пользователем BeGlobal Trainer:
  - Лингвист или переводчик
  - Имеющий опыт работы с MS Office и SDL Trados Studio
  - Имеющий представление о TMX, XLIFF и прочих специализированных форматах файлов
  - *Опыт работы с МП в целом и настройкой МП в частности не требуется*

Процесс обучения/создания модуля МП

Какие данные нужно подготовить для настройки

– Файлы в формате TMX (параллельные тексты в кодировке UTF-8). Можно загрузить в систему и целый архив файлов в формате TMX:

– Тестовый набор в формате TMX (Test Set), который будет использован для оценки по системе BLEU (опционально), не менее 100 сегментов.

- Если тестовый файл не загружен, система автоматически выберет случайную 1000 сегментов из доступных материалов для тестирования.

– Файл для регрессионного тестирования (Regression Test), формата TXT, в кодировке UTF-8, на исходном языке (опционально).

Как подготовить базу ТМ

1. **Проверить кодировку (UTF-8).**
2. **Удалить теги.**
3. **Удалить сегменты, в которых перевод идентичен исходному тексту.**
4. **Удалить некорректные сегменты (перевод не соответствует исходному тексту).**
5. **Никогда не удаляйте & > or < если они присутствуют в файлах.**
6. Дополнительные действия по проверке на:
   1. Некорректные символы
   2. Переносы внутри предложений
   3. Знаки пунктуации
   4. И т.д.

Как создать тестовый набор (Test Set)?

- Выберите наиболее репрезентативные для данного контента сегменты, но не менее 100 предложений.
- Данные предложения должны быть полными и грамматически корректными, не слишком короткими и не слишком длинными.
- Выбирайте предложения с различиями в стиле, терминологии и пр.
- Чем большее количество предложений выбирается для данной цели, тем более достоверными будут результаты оценки (**BLEU score**)
- Если вы не сможете создать тестовый набор самостоятельно, BeGlobal Trainer автоматически подготовит необходимую выборку, которую можно будет загружать в дальнейшем для тестирования соответствующих языковых пар.

Что такое BLEU Score?

- **BLEU** (**Bilingual Evaluation Understudy**) – это алгоритм для оценки качества текста после машинного перевода.
- Базовая посылка BLEU: «чем ближе результат МП к результату ручного перевода, тем он лучше”.
- Очки (scores) подсчитываются для конкретных переведенных сегментов (обычно предложений) путем сопоставления их с переводами хорошего качества. Затем результаты анализируются на всем корпусе текстов для оценки общего качества переводов. Понятность или грамматическая корректность во внимание не принимаются.
- Повышение BLEU score не является индикатором улучшения качества перевода.
- Результаты теста BLEU должны использоваться в ограниченных пределах, для сравнения результатов двух аналогичных систем, а также для отслеживания существенных изменений в рамках одной системы.

Как выбрать и использовать файл для регрессионного тестирования?

- Файл для регрессионного тестирования должен быть выбран на исходном языке для целей автоматического МП непосредственно после окончания настройки языковой пары
- Основные требования – те же, что и к тестовому набору, однако следует избегать того, чтобы предложения входили в состав базы, используемой для настройки, для получения слишком прекрасного результата.

```
1  Examine the work site for restricted traffic patterns, obstructed views, congestion, under-ground power or gas lines, etc.
2  We establish and adhere to environmentally sound policies and practices in product design, engineering, and manufacturing.
3  Solid mounted undercarriage standard on D3K through D5K provides stable platform for low impact, and high abrasion applications.
4  Well matched engine and power train systems enhance productivity and fuel efficiency.
5  Fewer parts and longer component life, less fuel and fluids, means less to replace and less to dispose.
6  The extended side wings, angled to thirty degrees, further increase the capacity over a standard blade.
7  The heavy-duty design includes reinforcement to transfer machine power without damaging the blade or the tractor.
8  Blade cylinders are pinned to the C-frame, and the blade height is such that the blade lift cylinders are isolated from damaging forces.
9  The taller blade allows pushing from a higher position, eliminating blade drag and increasing productivity.
10 Extended side plates make it easier for operators to "catch" the stinger when repositioning for a new pass.
11 The center of the blade is armored with T-1 plate steel for maximum service life.
12 With this design, coal-moving capacity can be as much as 200 percent greater than a standard U-blade.
13 A duty cycle of 4 percent with the engine off means that the sensor is working but sending a signal below the expected limit.
14 The Engine Idle Shutdown parameters define the shutdown strategy when the engine is operated at low idle for an extended period.
15
```

Normal text file | length : 1594 lines : 15 | Ln : 15 Col : 1 Sel : 0 | Dos\Windows | UTF-8 | INS

SDL BeGlobal - Mozilla Firefox

File Edit View History Bookmarks Tools Help

SDL BeGlobal

I Want To... Dashboard Trainer Translation Terms And Brands Reporting Account

TRAINER

Projects Language Pairs

* = Required

Language Pair Information

Project: ABC Corp

Name: * Web Articles

Source: * Spanish

Target: * English

File upload: Browse...

Note: Please select TMX training files containing parallel data (UTF-8 encoded).
One or more TMX files can be uploaded in a ZIP (.zip) file

Files Uploaded: 0 / 0

File Name	Size	Status	Progress
No files uploaded. Select browse to begin upload.			

Optional Settings

Submit Cancel

a1 demo5

SDL | BeGlobal™ ONLINE

SDL BeGlobal - Mozilla Firefox

File Edit View History Bookmarks Tools Help

SDL BeGlobal

I Want To... | Dashboard | Trainer | Translation | Terms And Brands | Reporting | Account

TRAINER

Projects | Language Pairs

* = Required

Language Pair Information

Project: ABC Corp

Name: * Web Articles

Source: * Spanish

Target: * English

File upload: * Browse...

Note: Please select TMX training files containing parallel data (UTF-8 encoded).
One or more TMX files can be uploaded in a ZIP (.zip) file

Files Uploaded: 1 / 2

	File Name	Size	Status	Progress	
✔	news.spa-eng.tmx	2.4 MB	Done	100%	✕
	news-less.spa-eng.zip	748.1 KB	Sending	26%	

Optional Settings

Submit | Cancel

a1 demo5

SDL | BeGlobal™ ONLINE

SDL BeGlobal - Mozilla Firefox

File Edit View History Bookmarks Tools Help

SDL BeGlobal

I Want To... | Dashboard | Trainer | Translation | Terms And Brands | Reporting | Account

TRAINER

Projects | Language Pairs

Viewing: ABC Corp | Delete | New Language Pair

0 of 3 Test Slots in Use

Being Trained (2)	Trained (1)	Being Tested (0)	Being Used (0)

TID: 2067 **News Articles**

SPA → ENG

Running 49%

(Est 9 hr)

Started: 08/27/2012 12:19 AM (GMT-07:00)

View & Manage

TID: 2069 **Web Articles**

SPA → ENG

Running 10%

(Est N/A)

Started: 08/27/2012 07:19 PM (GMT-07:00)

View & Manage

TID: 2039 **OpenOffice-UI**

SPA → ENG

Better

Started: 08/25/2012 09:08 PM (GMT-07:00)

Finished: 08/27/2012 04:24 AM (GMT-

Click here to view details and manage this training.

View & Manage

a1 demo5

SDL | BeGlobal™ ONLINE

SDL BeGlobal - Mozilla Firefox

File Edit View History Bookmarks Tools Help

SDL BeGlobal

I Want To... | Dashboard | Trainer | Translation | Terms And Brands | Reporting | Account

TRAINER

Projects | Language Pairs

< Back | Delete Training

	Training Results	Test	Ready to use
SPA → ENG	Better	LPID: N/A	
		Test	Activate
		0 of 3 Test Slots in Use	

Training Information

Training BLEU Score: **26.08**

Baseline BLEU Score: 22.63

Training ID: 2039

Name: OpenOffice-UI

User: a1 demo5

Language Pair: Spanish to English

LPID: N/A

Project: ABC Corp

Domain: Software

Description: N/A

Start Date: 08/25/2012 09:08 PM (GMT-07:00)

End Date: 08/27/2012 04:24 AM (GMT-07:00)

Elapsed Time: 1D:7H:15M:29S

a1 demo5

SDL | BeGlobal™ ONLINE

SDL BeGlobal - Mozilla Firefox

File Edit View History Bookmarks Tools Help

SDL BeGlobal

I Want To... | Dashboard | Trainer | Translation | Terms And Brands | Reporting | Account

TRAINER

Projects | Language Pairs

Viewing: ABC Corp | Delete | New Language Pair

1 of 3 Test Slots in Use

Being Trained (2)	Trained (0)	Being Tested (1)	Being Used (0)

TID: 2067 **News Articles**

SPA → ENG

Running 50%

(Est 9 hr)

Started: 08/27/2012 12:19 AM (GMT-07:00)

View & Manage

TID: 2069 **Web Articles**

SPA → ENG

Running 48%

(Est 9 hr)

Started: 08/27/2012 07:19 PM (GMT-07:00)

View & Manage

TID: 2039 **OpenOffice-UI**

SPA → ENG

Better

LPID: 2263

Started: 08/25/2012 09:08 PM (GMT-07:00)

Finished: 08/27/2012 04:24 AM (GMT-07:00)

View & Manage

a1 demo5

SDL | BeGlobal™ ONLINE

SDL BeGlobal - Mozilla Firefox

File Edit View History Bookmarks Tools Help

SDL BeGlobal

I Want To... Dashboard Trainer Translation Terms And Brands Reporting Account

TRAINER

Projects Language Pairs

< Back Delete Training

SPA → ENG

Training Information

Training BLEU Score: 26.

Baseline BLEU Score: 22.

Training ID: 203

Name: Op

User: a1

Language Pair: Spa

LPID: 226

Project: ABC

Domain: Sof

Description: N/A

Start Date: 08/25/2012 09:08 PM (GMT-07:00)

End Date: 08/27/2012 04:24 AM (GMT-07:00)

Elapsed Time: 1D:7H:15M:29S

dy to use

tivate

Evaluate Trained Language Pair

Instant Translation | Translate a Document

Select Language Pair: 2263: spa -> eng (OpenOffice-UI)

Si pone el cursor en la barra de título, mantiene pulsado el ratón sobre el símbolo y lo arrastra lejos de la barra, el menú desplegable seguirá al puntero del ratón.

Evaluation result: (LPID 2263)

If you place the cursor in the title bar, hold down the mouse over the icon and drag it away from the bar, the floating menu will follow the mouse pointer.

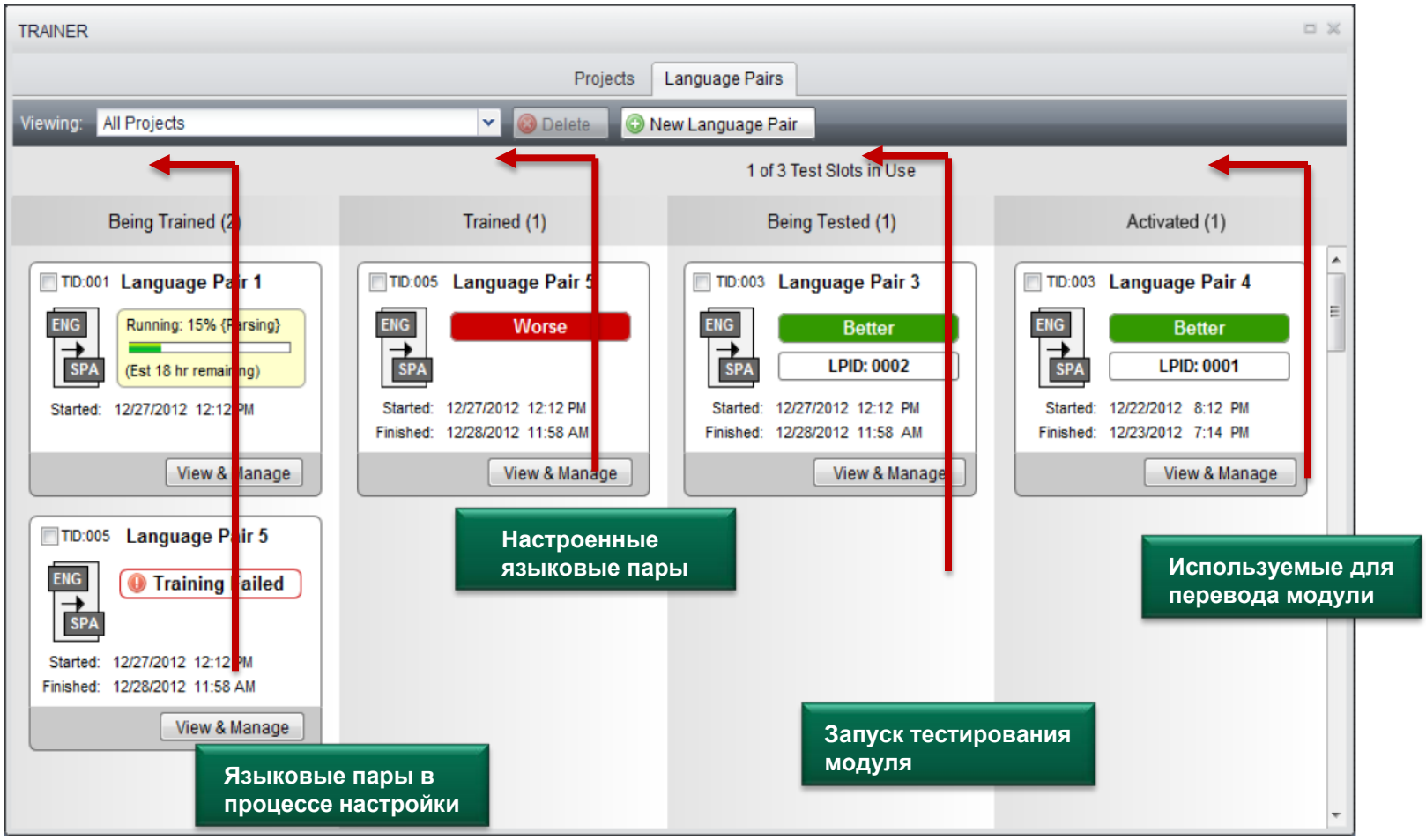
Baseline result: (LPID 1452)

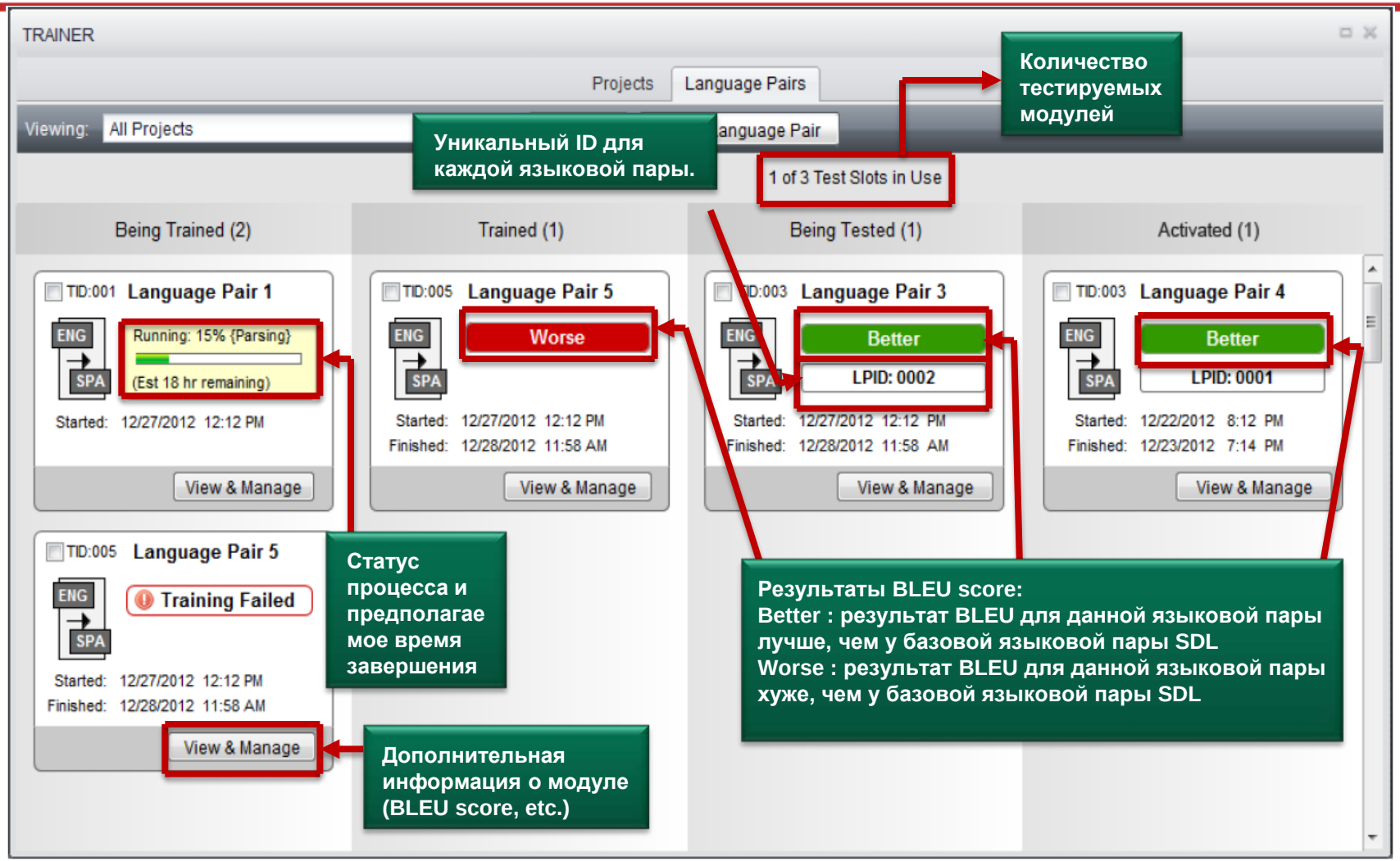
If you put the cursor in the title bar, holding down the mouse over the symbol and drag it away from the bar, the drop-down menu will follow the mouse pointer.

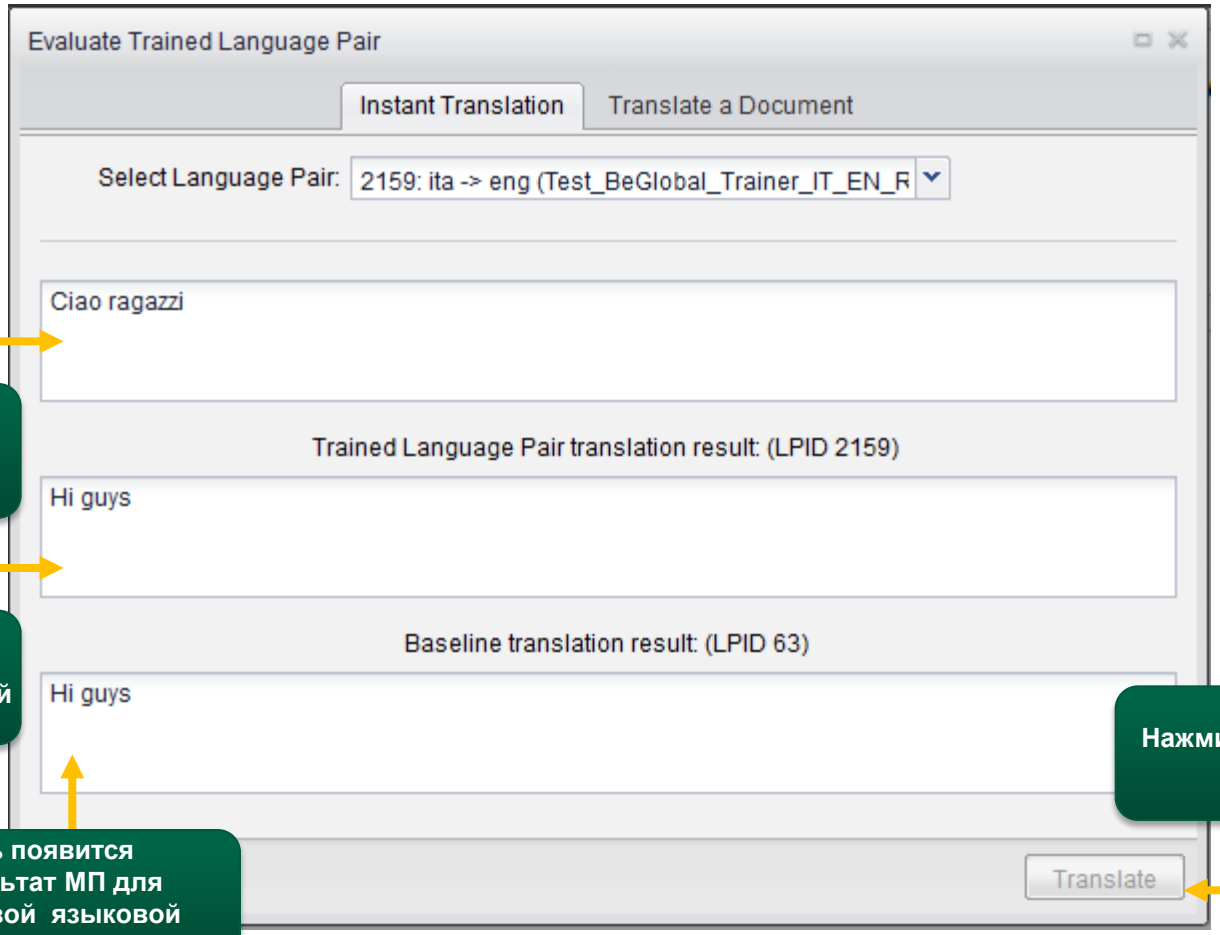
Translate

a1 demo5

SDL | BeGlobal


TRAINER
Projects
Language Pairs
Viewing: All Projects
Delete
New Language Pair
1 of 3 Test Slots in Use
Being Trained (2)
Trained (1)
Being Tested (1)
Activated (1)
TID:001 Language Pair 1
ENG
SPA
Running: 15% {Parsing}
(Est 18 hr remaining)
Started: 12/27/2012 12:12 PM
View & Manage
TID:005 Language Pair 5
Training Failed
Started: 12/27/2012 12:12 PM
Finished: 12/28/2012 11:58 AM
View & Manage
TID:005 Language Pair 5
Worse
Started: 12/27/2012 12:12 PM
Finished: 12/28/2012 11:58 AM
View & Manage
TID:003 Language Pair 3
Better
LPID: 0002
Started: 12/27/2012 12:12 PM
Finished: 12/28/2012 11:58 AM
View & Manage
TID:003 Language Pair 4
Better
LPID: 0001
Started: 12/22/2012 8:12 PM
Finished: 12/23/2012 7:14 PM
View & Manage
Настроенные языковые пары
Используемые для перевода модули
Языковые пары в процессе настройки
Запуск тестирования модуля


TRAINER
Projects
Language Pairs
Viewing: All Projects
anguage Pair
Уникальный ID для каждой языковой пары.
Количество тестируемых модулей
1 of 3 Test Slots in Use
Being Trained (2)
Trained (1)
Being Tested (1)
Activated (1)
TID:001 Language Pair 1
ENG
SPA
Running: 15% {Parsing}
(Est 18 hr remaining)
Started: 12/27/2012 12:12 PM
View & Manage
TID:005 Language Pair 5
Worse
Started: 12/27/2012 12:12 PM
Finished: 12/28/2012 11:58 AM
View & Manage
D:003 Language Pair 3
Better
LPID: 0002
Started: 12/27/2012 12:12 PM
Finished: 12/28/2012 11:58 AM
View & Manage
TID:003 Language Pair 4
Better
LPID: 0001
Started: 12/22/2012 8:12 PM
Finished: 12/23/2012 7:14 PM
View & Manage
TID:005 Language Pair 5
Training Failed
Started: 12/27/2012 12:12 PM
Finished: 12/28/2012 11:58 AM
View & Manage
Статус процесса и предполагае мое время завершения
Результаты BLEU score:
Better : результат BLEU для данной языковой пары лучше, чем у базовой языковой пары SDL
Worse : результат BLEU для данной языковой пары хуже, чем у базовой языковой пары SDL
Дополнительная информация о модуле (BLEU score, etc.)



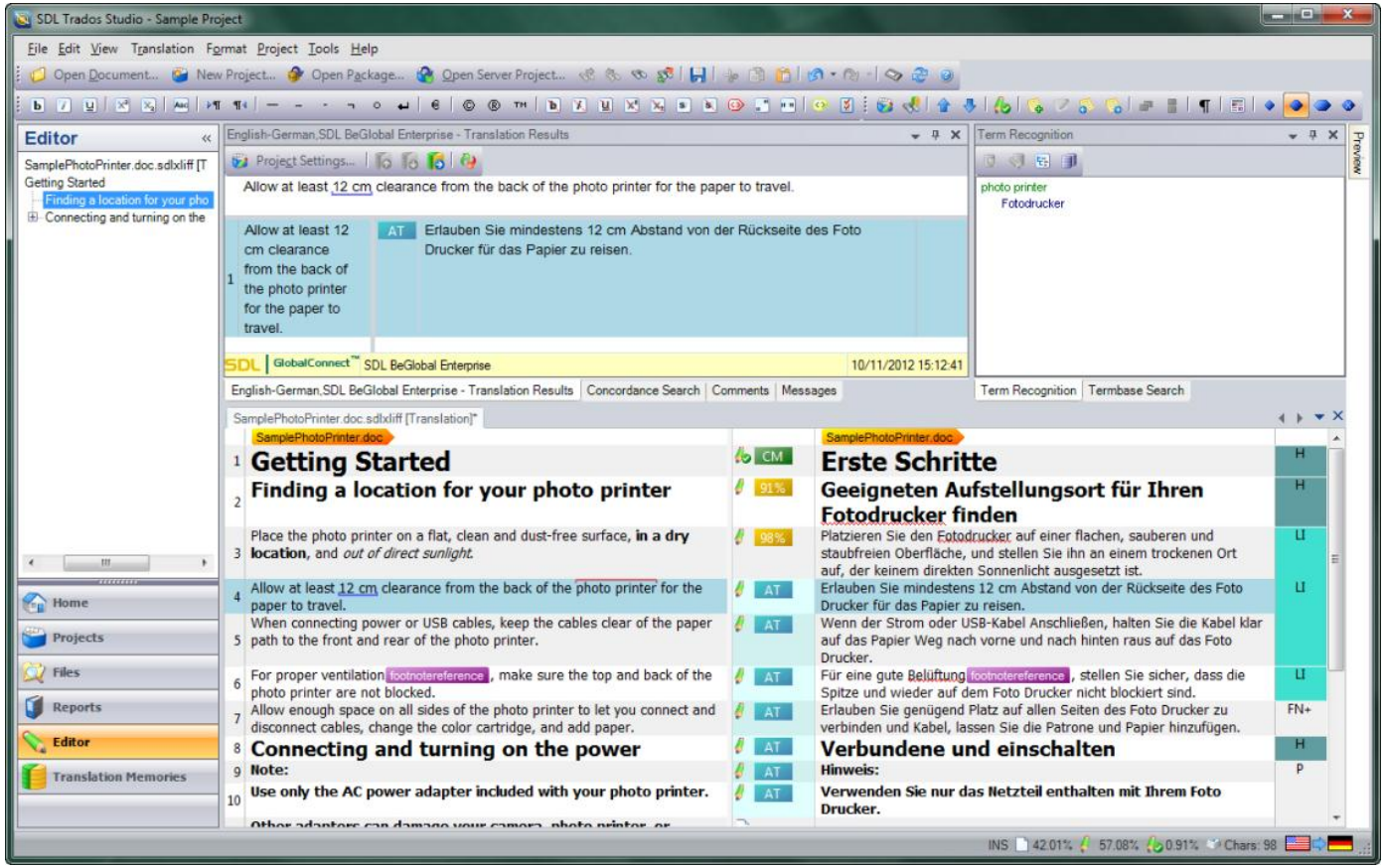
Здесь можно напечатать текст для перевода

Здесь появится результат МП для настроенной языковой пары

Здесь появится результат МП для Базовой языковой пары

Нажмите Translate.

Пример использования: машинный перевод как неполное совпадение в SDL Trados Studio



Глоссарий

SDL Proprietary and Confidential

- Activate - Активировать: перенести языковую пару на учетную запись для выполнения МП с помощью разных приложений.
- Baseline-Базовая: языковая пара, созданная SDL. Используется совместно с пользовательской, в том числе для поиска терминов.
- BLEU score: автоматически вычисляемая степень совпадения результата машинного и ручного перевода для одного и того же текста.
- Deploy-Развернуть: перенести языковую пару на тестовый сервер для выполнения переводов.
- Language Pair-Языковая пара: модуль МП и статистические данные, используемые для перевода
- LPID: уникальный номер создаваемой языковой пары.
- Match-Совпадение: соответствие между тестовым материалом и текстовым корпусом, используемым для создания языковой пары. В системе САТ обычно понимается как ТМ-совпадение.
- N-gram: для целей СМП термин означает последовательное количество (“n”) слов, которые появляются в отдельно взятом сегменте. Например, “bigram” означает комбинацию из 2 слов; 4-gram – 4 последовательных слова, и т.д.

- TMX files-Файлы в формате TMX: файлы в формате Translation Memory eXchange, которые требуются для создания обученной языковой пары.
- Parallel data-Параллельные данные: сопоставленные предложения на исходном и выходном языках.
- Regression Testing-Регрессионное тестирование: документ для автоматического перевода после настройки языковой пары с целью сравнения качества МП, полученного с помощью различных модулей (базовой системы и настроенной).
- Source language-Исходный язык: язык документа.
- Target language-Выходной язык: язык перевода.
- Test Set-Тестовый набор: файл для получения результата BLEU score.
- Test slot-Тестовый слот: место на сервере, где имитируется процесс активации языковой пары для использования в BeGlobal с целью оценки качества.
- Train/training-Обучение: процесс сопоставления исходных и выходных пользовательских данных из файлов TMX в модуле статистического МП с целью создания новой языковой пары на основе математических алгоритмов
- Un-deploy-Отменить развертывание: удалить языковую пару с тестового сервера для освобождения тестового слота, при этом языковая пара физически не удаляется.

Продолжение дискуссии:
группа UTIConf