SDL Your Content Their Language

# Технологии и системы машинного перевода от компании SDL

SDL

# Общие сведения о машинном переводе

- **Машинный перевод (МП)** — перевод текста в цифровом формате с одного естественного языка на другой с помощью компьютера
- **Главной проблемой,** с которой сталкиваются системы машинного перевода, является **совокупная неоднозначность всех слов и всех грамматических правил**

*“Я ел суши с тунцом" — можно истолковать по-разному, в том числе и как: "Тунец составил мне компанию за столом".*

- **Задача системы машинного перевода** состоит в принятии грамотных решений относительно выбора значений слов и применения грамматики для перевода
- **Методика,** применяемая в системе для устранения неоднозначности и принятия решений, зависит от подхода, использованного в процессе ее разработки (основные типы систем будут рассмотрены далее)
- **Два подхода к решению проблемы**: в течение последних десятилетий неоднократно предпринимались попытки решить проблему неоднозначности и повысить качество результатов машинного перевода. Было доказано и опровергнуто множество теорий, что в итоге привело к появлению двух основных подходов**: машинному переводу на основе правил и статистическому машинному переводу,** а также **к гибриду** на основании этих двух типов

# Основные типы систем машинного перевода

- **Машинный перевод на основе правил (Rule-Based Machine Translation, RBMT)**

*Компания Common Sense Advisory (www.commonsenseadvisory.com) определяет машинный перевод на основе правил следующим образом: "Перевод, основанный на лингвистических теориях, грамматических структурах или значениях. Лингвисты и программисты кодируют эту информацию в виде правил и баз данных".*

- Лингвистические правила, созданные вручную
- Словарные базы данных, созданные как вручную, так и автоматически
- Отсутствие к.-либо правил и словарей > качество перевода
- Кроме того, существует множество правил о применении правил в определенном контексте
- По прошествии некоторого времени повышать качество перевода все сложнее, так как новые правила, созданные для определенных сценариев, могут отрицательно повлиять на качество перевода с использованием других сценариев
- В итоге качество работы с языковыми парами в таких системах в дальнейшем повышается незначительно
- Для разработки новых языковых пар, даже на базовом уровне, требуется от нескольких месяцев до нескольких лет, так как необходимо создать большое количество правил вручную; для написания правил требуется группа специалистов, обладающих глубоким знанием исходного и целевого языков
- Для настройки системы под заказчика ограничивают область ее применения конкретным пользователем или предметной областью. В ходе процесса анализируются данные, полученные от клиента, и создаются соответствующие правила и словарные статьи. Эта процедура выполняется в значительной степени вручную - и требует временных и трудовых затрат. Кроме того, как другие процедуры, выполняемые вручную, она является дорогостоящей

# Основные типы систем машинного перевода

## Статистический машинный перевод (Statistical Machine Translation, SMT)

*Статистический машинный перевод определяется Common Sense Advisory следующим образом: "Вероятностный подход с применением статистического анализа паттернов, встречающихся в больших объемах текстов. Проще говоря, текстовые данные подвергаются компьютерному анализу, разбиваются на символьные строки, в которых определяются паттерны, затем эти закономерности используются при переводе"*

- Процесс создания языковых пар включает в себя обработку и изучение системой предварительно переведенных текстов, а также самообучение применению правил для устранения неоднозначности при переводе с одного языка на другой; значительная часть разработки полностью автоматизирована
- Кроме того, система обучается применению правил в определенном контексте
- После создания и запуска в работу языковой пары точность перевода можно повысить с помощью ввода новых данных в процесс обучения и совершенствования статистических алгоритмов для процесса обучения и перевода
- Чем больше доступных для автоматической обработки данных, тем выше качество перевода
- **Системы обоих основных типов успешно используются в настоящее время**, однако с увеличением объема информации для перевода с различных языков становится очевидным, что статистический машинный перевод обладает рядом преимуществ, в том числе:

- скоростью разработки новых языковых пар (месяцы по сравнению с годами) – путем автоматической обработки предварительно переведенной документации для конкретной языковой пары
- простотой разработки специализированных версий системы для конкретных предметных областей или клиентов (IT, медицина, производство); процесс самообучения повышает скорость настройки
- сокращением срока обновления системы и повышения качества перевода (недели по сравнению с месяцами и годами)

# Основные типы систем машинного перевода

- **Гибридные системы машинного перевода (Hybrid Machine Translation)**
  - Автоматическое обучение + "грамматика" на основе правил

- Хотя подход на основе правил и статистический подход исторически рассматривались отдельно друг от друга, в настоящее время разработчики с обеих сторон работают над гибридными системами, объединяющими оба механизма для создания "лучшего в своем роде" инструмента для перевода
- Способы достижения цели тех и других отличаются друг от друга и используют преимущества существующих технологий.
- Разработчики систем машинного перевода на основе правил внедрили в качестве дополнительного элемента статистический анализ, чтобы повысить связность перевода
- Статистический синтаксис (разработка компании Language Weaver) объединяет ключевое преимущество статистического подхода — самообучение, с преимуществом подхода на основе правил — грамматикой. Эта гибридная версия статистической системы машинного перевода не только распознает грамматику, но и автоматически обучается правилам ее применения. В результате точность переводов на основе статистического подхода повышается, благодаря усовершенствованным алгоритмам обучения.

# Резюме:
# основные типы машинного перевода

- **Машинный перевод на основе правил (RBMT)**
  - Лингвистические правила
  - Словарные базы данных
  - > отсутствие правил > качество перевода
- **Статистический машинный перевод (SMT)**
  - Использование вероятностного подхода (на анализе и паттернах)
  - Изучение предварительно переведенных данных
  - Способность к статистическому обучению
  - > ввод данных> качество перевода… способность к обучению
- **Гибридные системы машинного перевода (HMT)**
  - Технологии объединены в новом подходе к машинному переводу — "статистическом синтаксисе"
  - Автоматическое обучение + "грамматика" на основе правил
  - Система распознает "грамматику" и способна к автоматическому обучению

# Базовые и специализированные версии коммерческих систем МП

- В системах машинного перевода, доступных на рынке, обычно представлены два варианта версий: **базовые и специализированные**

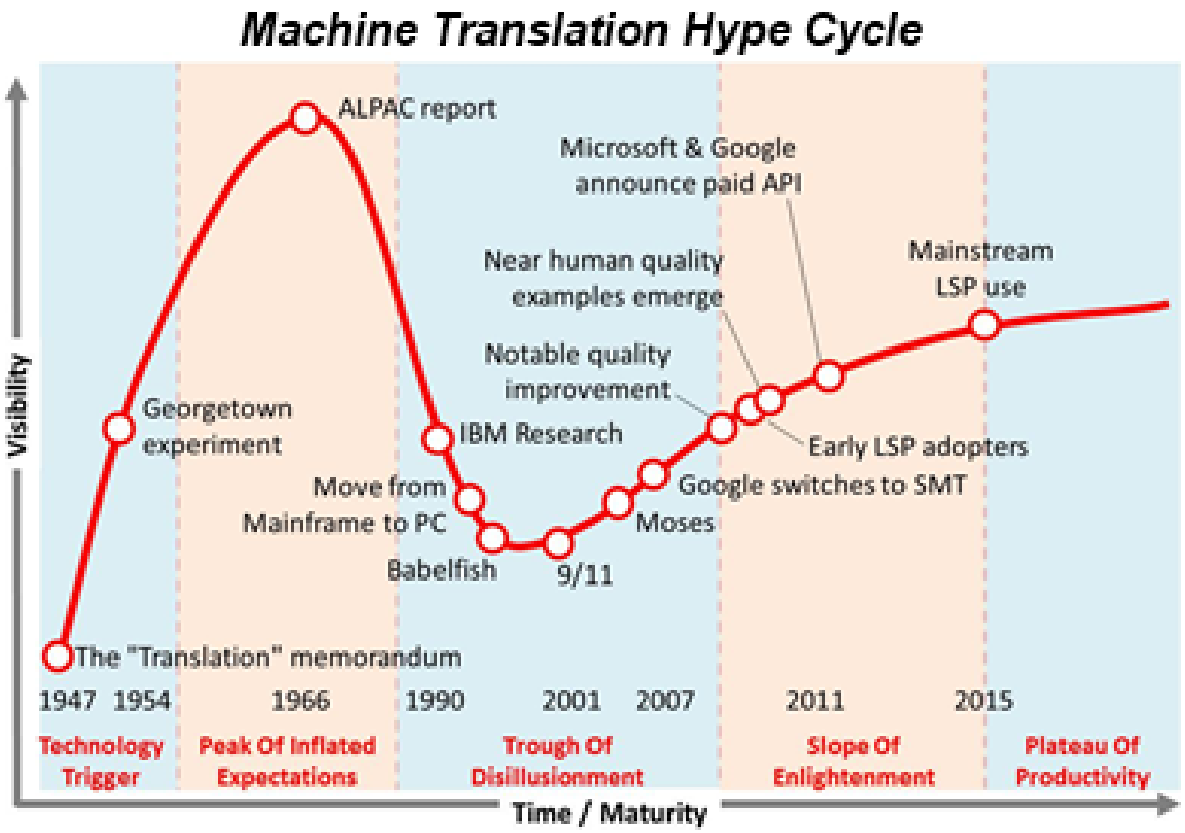
- **Базовая версия** входит в стандартный комплект программы.

Обычно такие системы разрабатывают для смыслового перевода различных тематик, системы предназначены для широкого применения

- С другой стороны, разработка **специализированной языковой версии** требует дополнительных усилий по созданию высококачественного перевода для определенного заказчика или конкретной предметной области, особое внимание уделяется терминологии и стилю

- В современных системах МП предлагаются мощные средства для настройки силами самих пользователей

# Gartner’s Hype Cycle для МП



# SDL и МП

## 10-летний опыт

### Интегрированные решения

SDL Intelligent Machine Translation (iMT)
для эффективного постредактирования МП

SDL BeGlobal
для полностью автоматизированного перевода (FAUT)

### Выделенные ресурсы

Постредакторы SDL

Лингвисты SDL

Разработчики МП

### Лидирующая технология МП

SDL Language Weaver (СМП)

Гибкая облачная инфраструктура

# Краткая история МП в SDL

- 2000: приобретение модуля RBMT у компании Transparent  Language, создание SDL Enterprise Translation Server (ETS)
- 2009: партнерство с компанией Language Weaver и изучение процессов настройки SMT-системы
- Более 7 млрд.слов, переведенных с помощью МП в SDL
- 2010: приобретение компании Language Weaver и продолжение работ над SMT
- 2013: новые возможности настройки МП: выпуск BeGlobal Trainer

# Компания SDL – о машинном переводе

- **Марк Ланкастер, основатель и исполнительный директор компании SDL (Великобритания):**
  - "В течение следующих пяти лет ожидается десятикратное увеличение объема цифровых данных"
  - "Мы считаем, что машинный перевод станет неотъемлемой частью процесса создания информационных ресурсов и стратегии управления компаний"
  - "Ожидается, что в течение последующих 5 лет свыше 30% информационных ресурсов будут переводиться машинным способом"
  - "Мы считаем Language Weaver самой передовой технологией машинного перевода, доступной на мировом уровне на данный момент"
  - "Интеграция защищенной и специализированной технологии машинного перевода в пакет технологий управления глобальной информацией SDL (GIM) обеспечивает возможность поддержки клиентов при создании глобальных информационных ресурсов в будущем".

SDL | BeGlobal™

SDL | BeGlobal™

- Облачная платформа
- Решение промышленного уровня
- Безопасность
- Специализация
- Надежные, качественные результаты
- Ориентация на разных пользователей
- Беспрецедентный уровень контроля

# SDL BeGlobal

- **Первая облачная платформа для машинного перевода в реальном времени**

- **Возможность привлечения пользователей к улучшению качества перевода**

Это первый продукт промышленного класса, в котором для повышения качества перевода используется привлечение пользователей к процессу. Теперь все пользователи могут предлагать лучшие варианты перевода выбранных ресурсов. Возможность привлечения пользователей к переводу в SDL BeGlobal может способствовать повышению качества переводимых информационных ресурсов в короткие сроки, увеличить степень текущего привлечения клиентов и снизить затраты на перевод и редактирование

- **Гибкий доступ к процессу перевода с помощью приложения рабочего стола**

С помощью нового виджета перевода для рабочего стола пользователи могут применять возможности платформы SDL BeGlobal на своих компьютерах. Простым перетаскиванием документов и ссылок URL в окно виджета обеспечивается доступ к защищенному корпоративному приложению. Таким образом риск появления секретной информации компании на открытых незащищенных сервисах для перевода снижается

- **Новые возможности использования SDL TrustScore**

Алгоритм SDL TrustScore, выигравший несколько наград и заявленный на патент, позволяет получить оценку качества каждого перевода. Кроме оценки качества на уровне документа, TrustScore работает на уровне сегментов, что упрощает для переводчиков и пользователей процесс принятия решений и редактирования

- **Дополнительные настройки безопасности**

Новые настройки безопасности в системе SDL BeGlobal обеспечивают физическое разделение данных на протяжении всего жизненного цикла перевода. Разделение данных позволяет согласовать процесс перевода с регулирующими процедурами безопасности компании и отдела IT и выполнить дополнительные требования по обеспечению безопасности.

# SDL BeGlobal

- **TouchPoints™** позволяет корпоративным пользователям определять и контролировать правильное содержимое, язык, качество время и место осуществления брендовых коммуникаций
- Надежные результаты. **TrustScore™** позволяет оценить качество каждого перевода. Пользователи сами закладывают ожидаемый уровень качества
- **Управление терминологией и названиями брендов** позволяет пользователям определить способ перевода названий торговых марок и основной терминологии
- Механизмы автоматического повышения качества перевода непрерывно совершенствуют последующие переводы
- **Виджеты** для привлечения пользователей к переводу позволяют получать обратную связь и предложения по более правильному переводу от сообщества пользователей (НОВИНКА)
- Удобство использования
- Централизованный пункт управления всеми коммуникациями в реальном времени
- Интуитивно понятный пользовательский интерфейс потребительского уровня, доступный каждому корпоративному клиенту
- Встроенные функции отчетов и аналитики
- Настройки гибкого доступа
- Веб-приложение не требует установки. Просто введите адрес и начинайте работу
- Программный интерфейс **REST API** обеспечивает простую интеграцию с существующей инфраструктурой
- Прединтеграция с другими системами с помощью надстройки **SDL GlobalConnect**
- Простой в использовании **виджет рабочего стола**, обеспечивающий доступ к защищенному переводу всем сотрудникам компании (НОВИНКА)
- Надежность и безопасность
- Многопользовательский доступ к приложению как услуге совмещен с ролевой моделью прав доступа, что позволяет обеспечить разные уровни безопасности
- Новые настройки развертывания с высокой степенью безопасности по требованию (НОВИНКА)

# SDL BeGlobal – это SMT-система

- Система “учится”, как переводить, анализируя корпус текстов
- Все базовые системы создаются на огромных массивах данных
- Широкие возможности настройки оптимизируют МП для целей конкретных предметных областей
- При создании новой языковой пары SMT требуется гораздо меньше усилий по программной разработке и настройке, чем RBMT

# Состав SDL BeGlobal

**BeGlobal Trainer**
Приложение для настройки МП

**BeGlobal Online**
Администрирование, терминология, отчеты…

**BeGlobal REST API**

**BeGlobal**
Облачная платформа для автоматического перевода

**Приложения**
BeGlobal Translator
BeGlobal for Microsoft Office
BeGlobal Desktop
BeGlobal for Websites

**Плагины**
RightNow Answers
RightNow Incidents
LivePerson Chat
Salesforce Knowledge

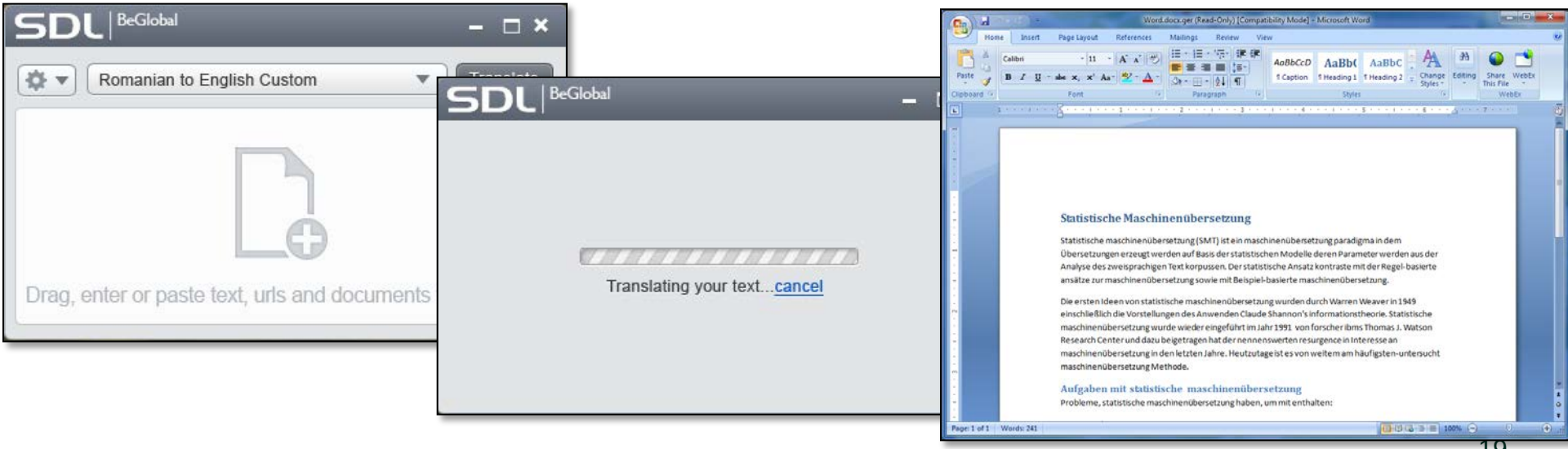
**Интеграция**
SDL Trados Studio
SDL WorldServer
SDL TMS

# BeGlobal vs BeGlobal Community Edition

- **What's the difference between this and the Community Edition?**
  - *The Community Edition is not trainable, does not include reporting on volume throughput, user management, terminology or additional options around security. It is however, ideal for anyone wishing to look at the quality of the baselines before investment.*

# Пример приложения для рабочего стола - SDL BeGlobal Desktop

- Инструмент для перевода текстов, файлов и веб-сайтов в безопасном режиме напрямую с рабочего стола.
- Загружается через SDL BeGlobal и запускается в операционной системе Windows или Mac OS X.
- Преимущества:
- Сотрудники могут выполнять защищенные переводы, используя специализированные языковые пары, списки терминов и применяемую процедуру отчетности без регистрации в веб-приложении.
- Перетаскивайте документы, вводите текст, переводите сайты.



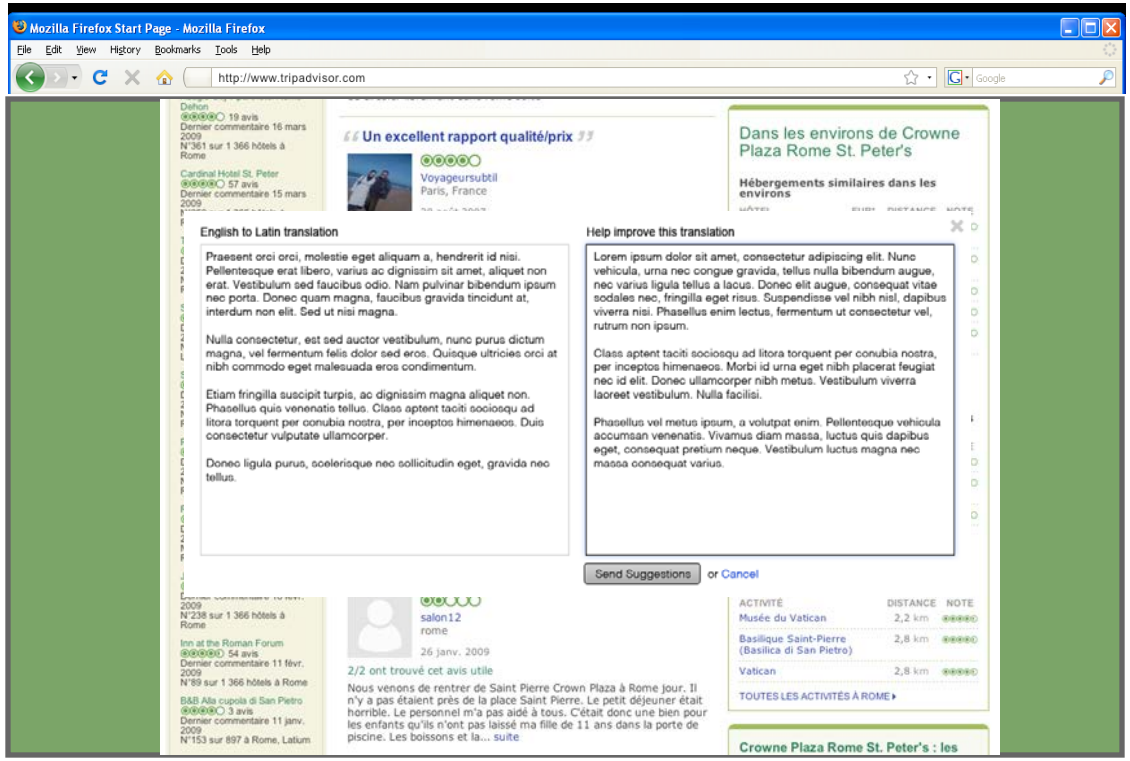
# Пример приложения: виджет предложений по переводу

**Что это такое?**

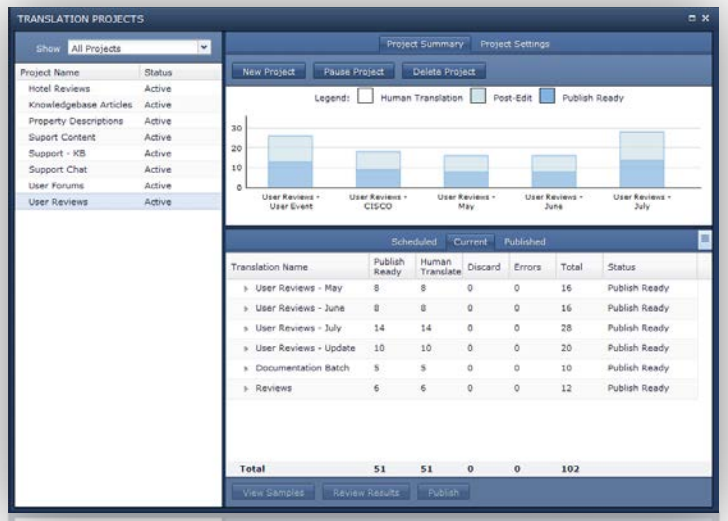
- **Виджет для получения от пользователей оценки и предложений по более правильному переводу.**

**Преимущества:**

- **Повышение качества языковых пар с помощью повторного обучения и предложений пользователей, а также снижение затрат по сравнению с обычным постредактированием.**



# Компоненты SDL BeGlobal




TouchPoints™

TrustScore™

Управление словарями терминов и брендов

Программный интерфейс REST API

Аналитика и отчеты

# TouchPoint™

- Простая и удобная установка
- Мощность, пользовательские настройки
- Позволяет определять и контролировать тип ресурсов, языки, качество и время и, например, место осуществления многоязычных коммуникаций компании с клиентами, будь то двусторонняя дискуссия потребителей на форуме компании или многоязычный чат представителя отдела продажи и потенциального покупателя.

Языковая пара

Термины и бренды

TrustScore™

Отчеты

# TrustScore™

- **Что такое TrustScore?**
  - Прогнозируемый критерий качества в диапазоне от 1 до 5
  - Калибровка в соответствии с требованиями компании

- **Преимущества TrustScore™**
  - Публикация только достоверных данных на любом языке
  - Определение способа управления переводом (отклонение, рецензирование или публикация)

# Исследование качества: TrustScore™

- **TrustScore™: Стратегический алгоритм для определения возможной оценки качества перевода человеком**

**TrustScore™** предоставляет владельцам ресурсов надежный способ публикации только тех переводов, которые в наибольшей степени соответствуют ожиданиям конечного пользователя, действуя на основании цели документа или коммуникации

# TrustScore™

## Принцип действия

- Автоматическая генерация алгоритмом при считывании каждого перевода
- Алгоритм,настроенный экспертами, выбирается клиентом для представления точки зрения конечного пользователя
- Экспертная оценка и калибровка выполняются в течение первых 4 недель процесса обучения

## Применение TrustScore™

- В настоящее время TrustScore™ применяется во всех модулях SDL BeGlobal

.

# Уровни оценки качества

5 Actionable and consistent with brand voice

4 Actionable with some brand voice (some errors)

3 Actionable

2 May be actionable

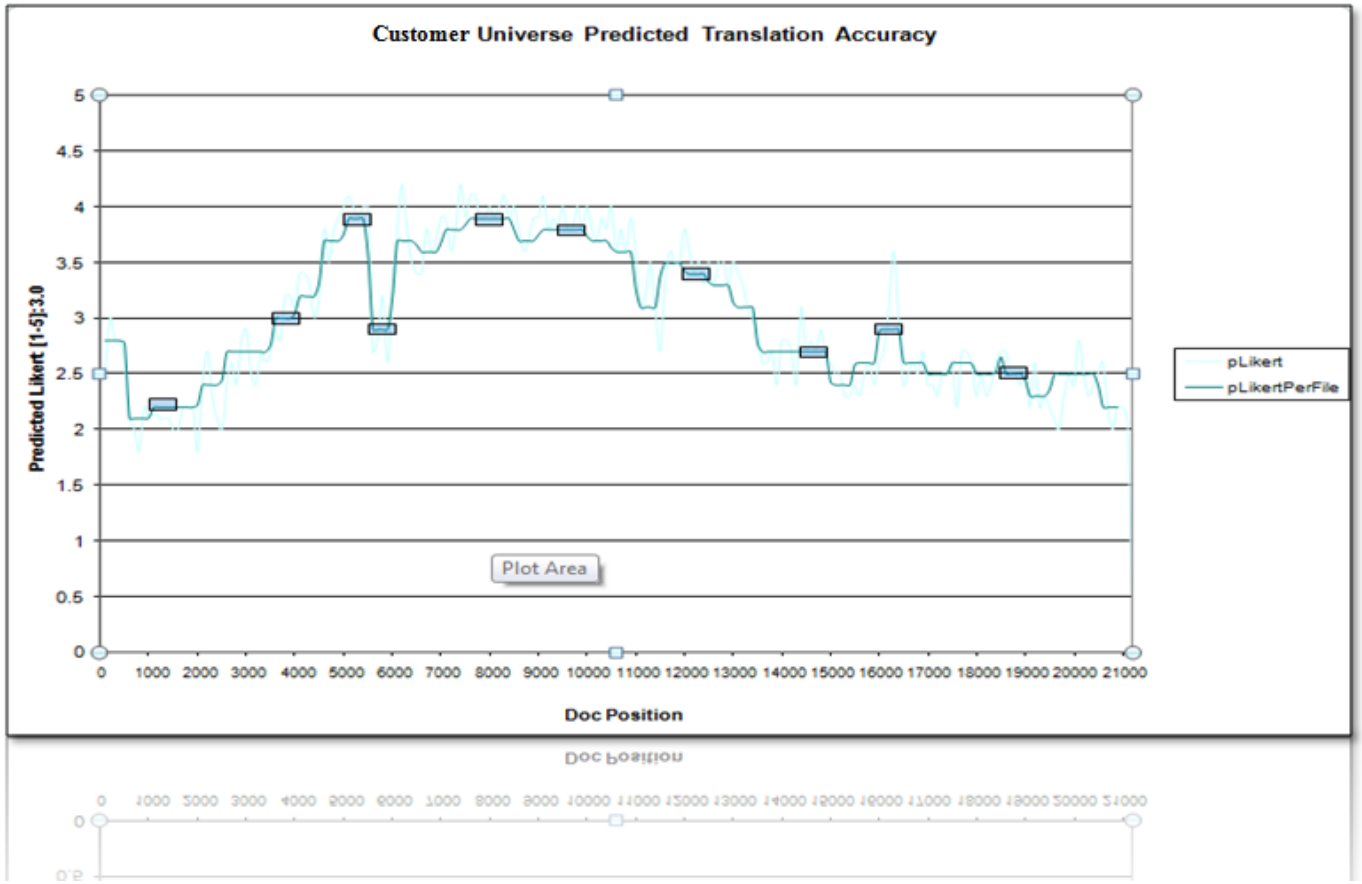
1 Not useful

Actionable Content

+ Brand Voice

# Автоматический прогноз качества

**Также выполняется автоматический прогноз качества перевода для каждого документа, затем документы группируются в зависимости от уровня качества.**

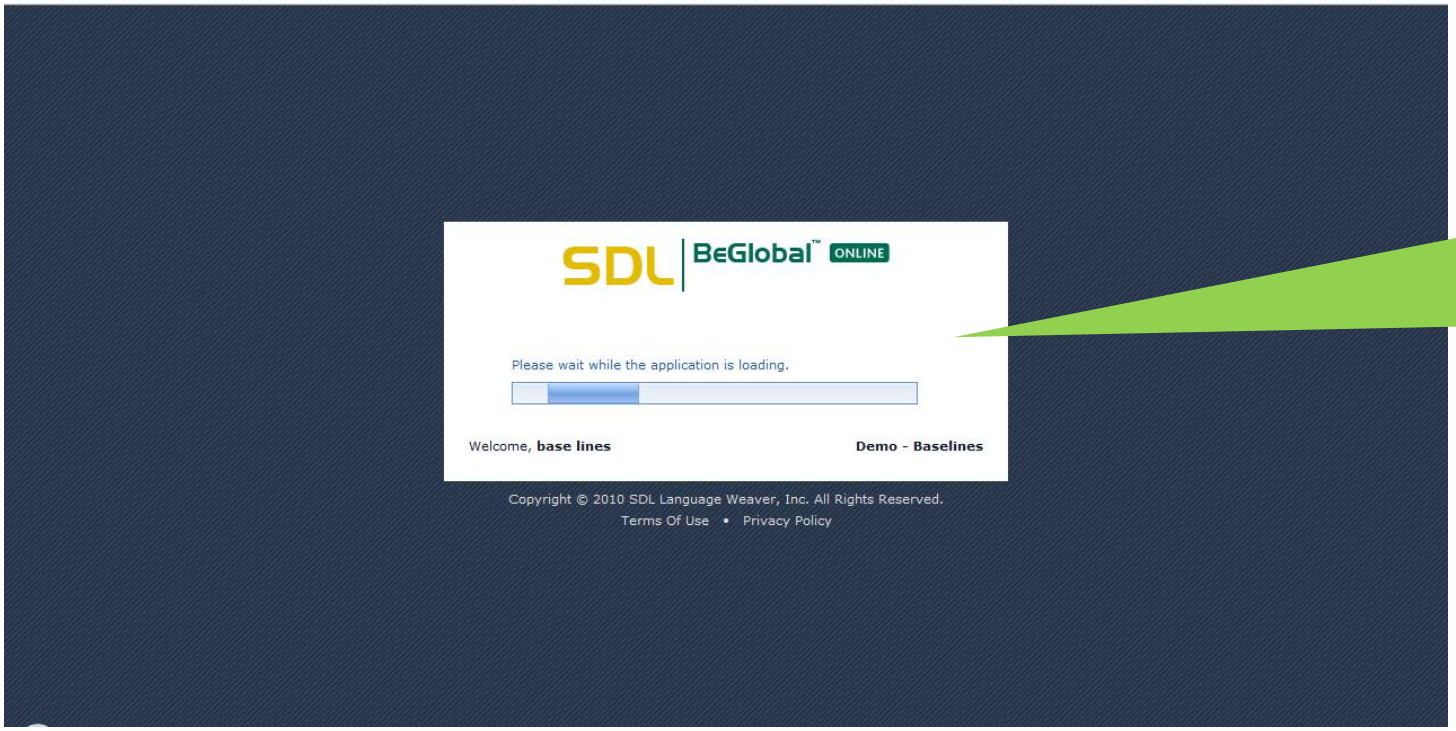


# SDL BeGlobal: вход в систему

***Шаг 1:***
зайдите на страницу регистрации SDL BeGlobal

***Шаг 2:***
введите адрес электронной почты и пароль

***Шаг 3:***
согласитесь на загрузку приложения




В диалоговом окне входа в систему будут загружены ваши настройки для доступа к переводам

# SDL BeGlobal: рабочая область

**Шаг 1:**

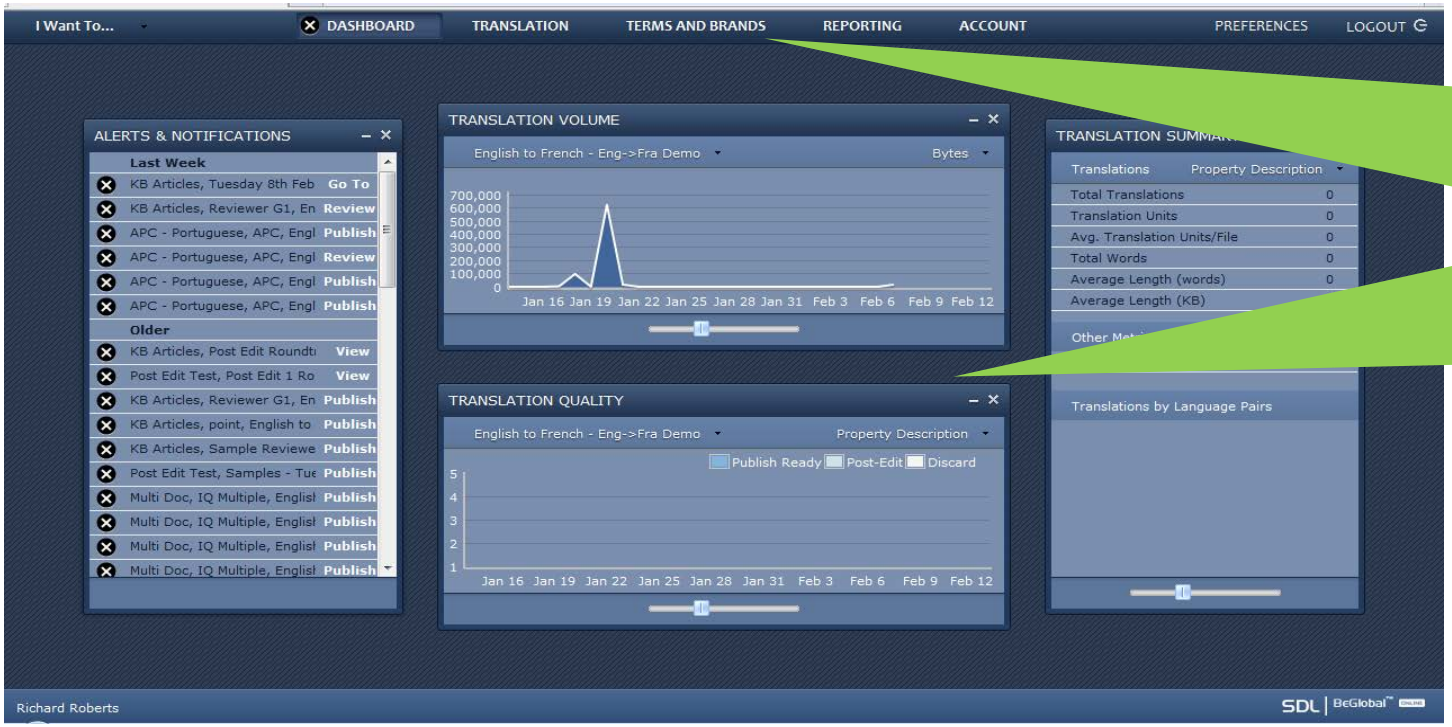
загрузка рабочей области

**Шаг 2:**

отображение 4 основных областей просмотра перевода

**Шаг 3:**

использование двух типов меню




Получить доступ к опциям меню можно через меню верхнего уровня или... команду "I want to..." ("Я хочу")

# SDL BeGlobal : словари терминов

**Шаг 1:**

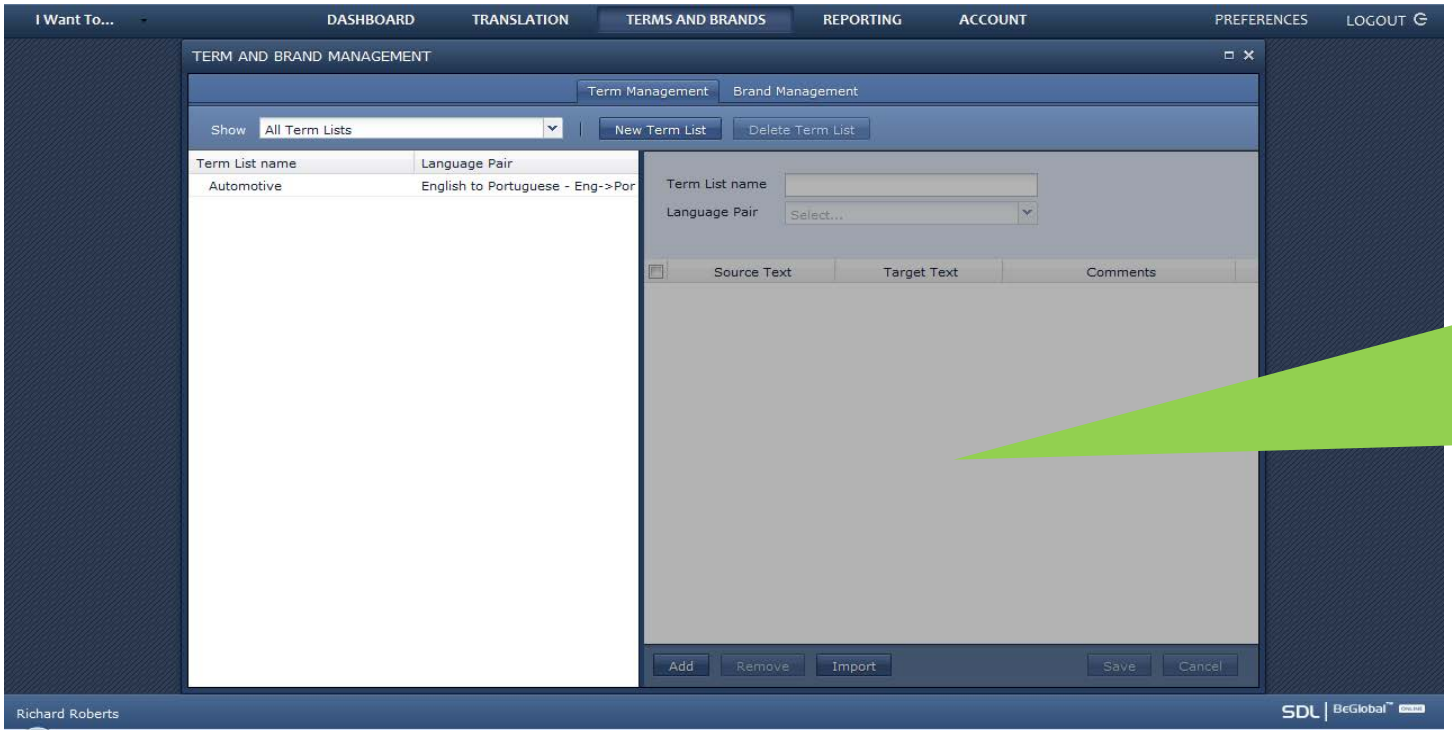
словарь терминов и брендов

**Шаг 2:**

пользователи могут обновлять оба словаря

**Шаг 3:**

добавление новых терминов




Словари терминов и названий брендов позволяют добавлять новые записи для повышения качества перевода

# SDL BeGlobal : словари брендов

**Шаг 1:**

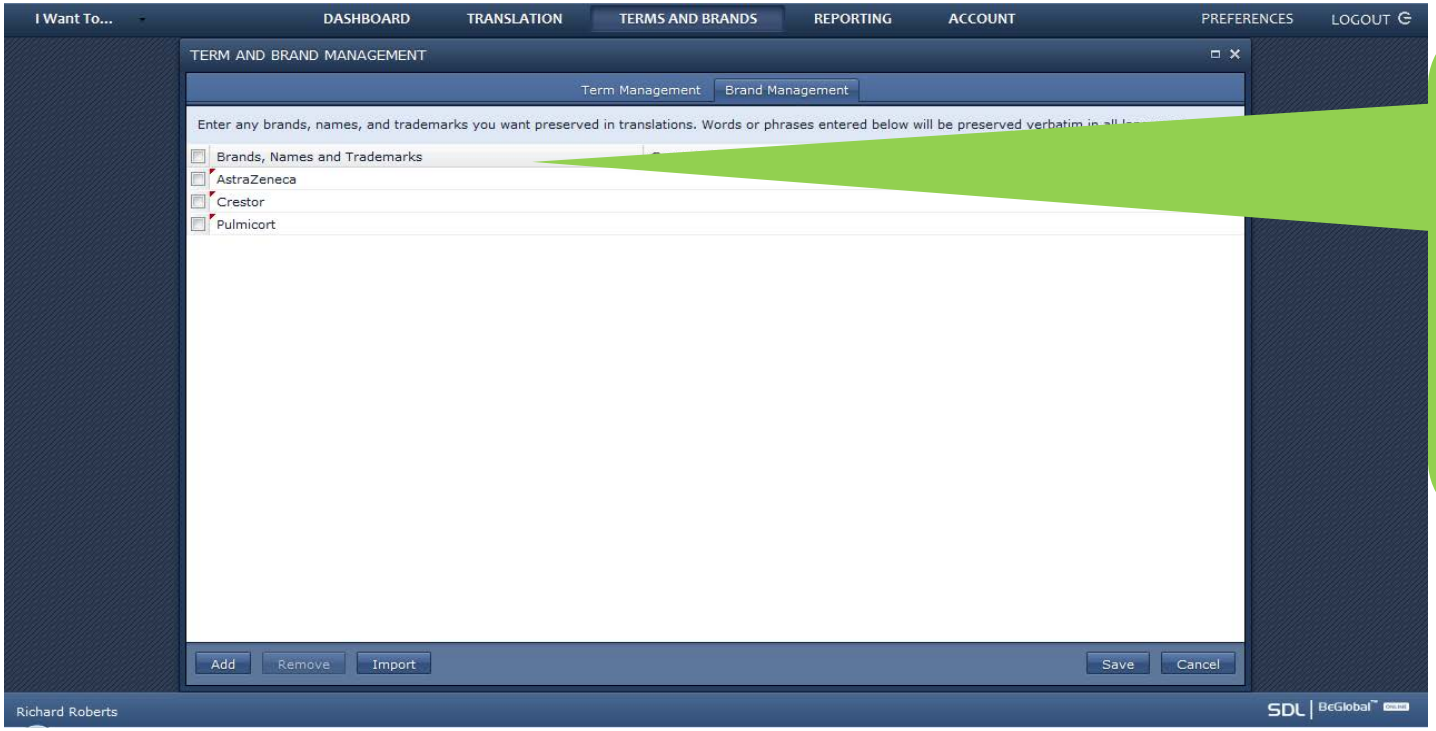
добавление записей в словарь брендов

**Шаг 2:**

выберите "Add" (Добавить"), затем сохраните

**Шаг 3:**

записи будут применимы ко всем переводам




Список брендов может включать в себя названия компаний, наименования продуктов, места расположения и т.д. Также можно добавлять комментарии

# SDL BeGlobal : Touchpoint (1)

***Шаг 1:***

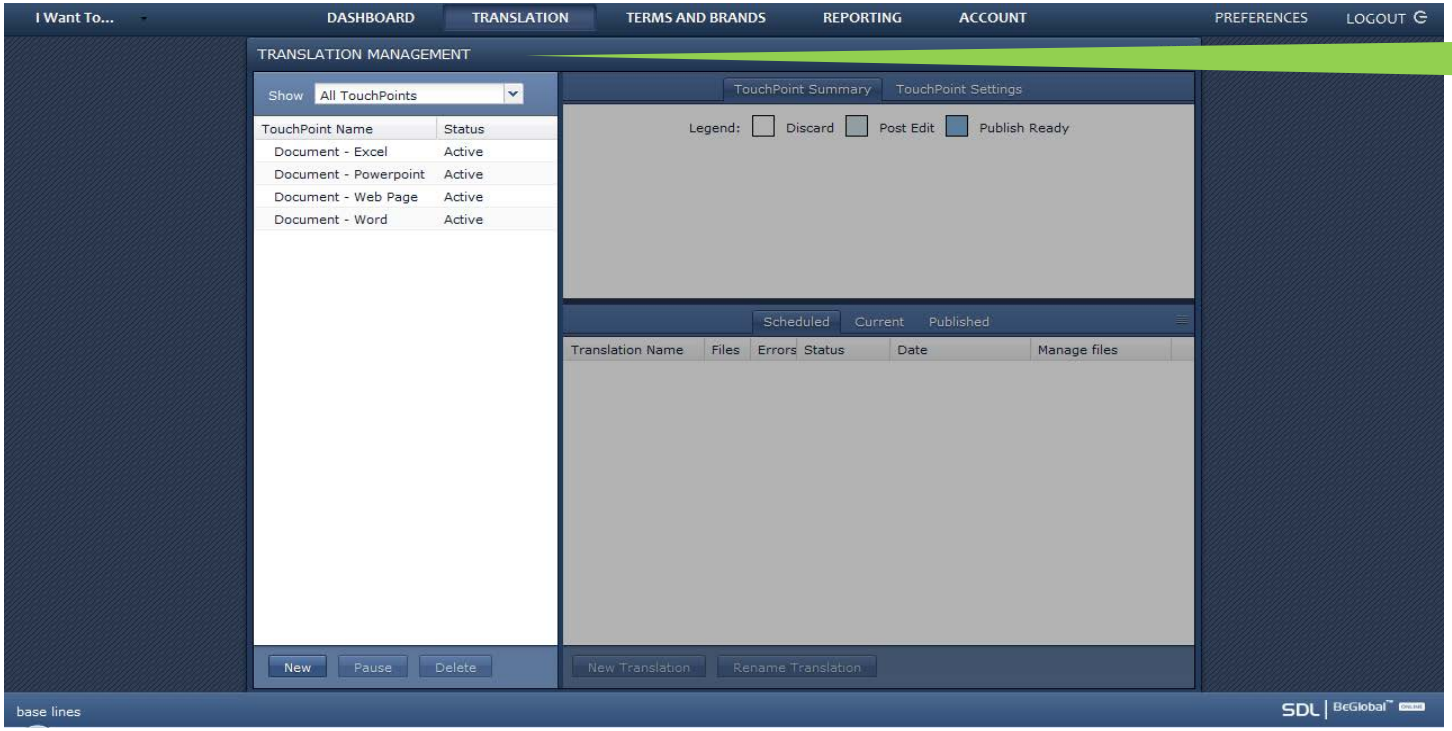
Touchpoint – основа решения

***Шаг 2:***

выберите из списка нужный тип перевода

***Шаг 3:***

после выбора будут загружены параметры перевода




Параметры перевода

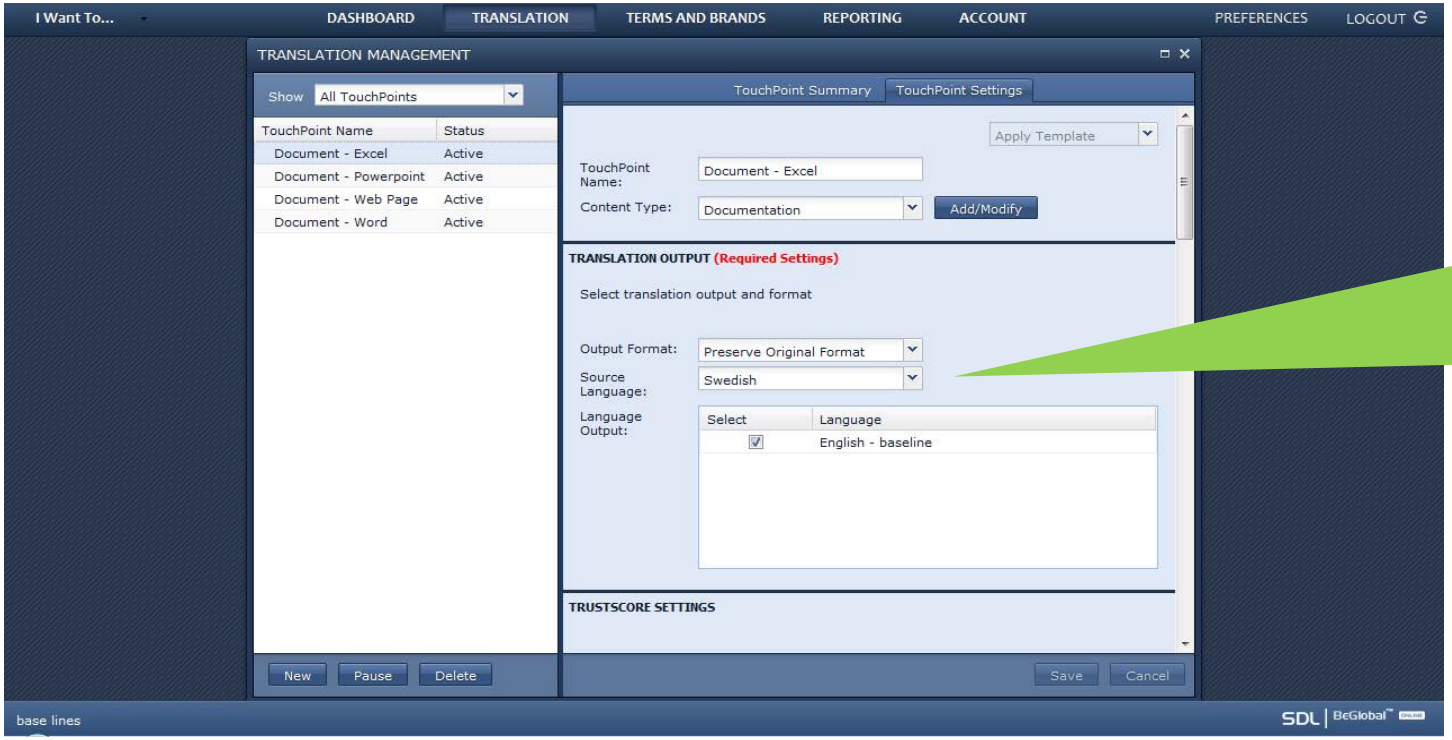
# SDL BeGlobal : Touchpoint (2)

***Шаг 1:***

после выбора появятся настройки для перевода

***Шаг 2:***

выберите вкладку "Touchpoint Summary" ("Сводка")



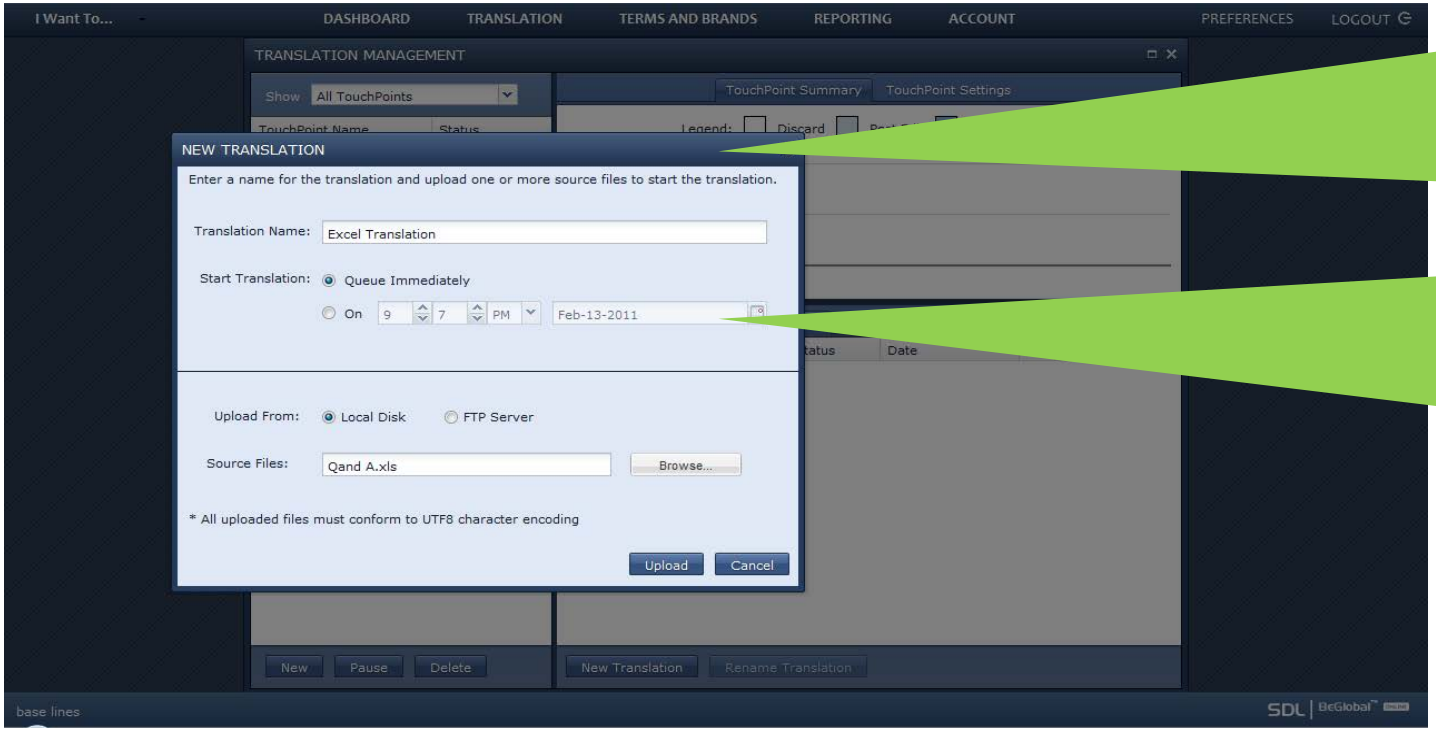

В настройках отображаются документы, выбранные для перевода со шведского на английский язык

# SDL BeGlobal : Touchpoint (3)

***Шаг 1:***
выберите New Translation ("Новый перевод") для вызова диалогового окна

***Шаг 2:***
введите название файла перевода

***Шаг 3:***
выберите файл для перевода




В диалоговом окне появится запрос на выбор названия для файла перевода Выберите в обозревателе файл для перевода

# SDL BeGlobal : Touchpoint (4)

***Шаг 1:***

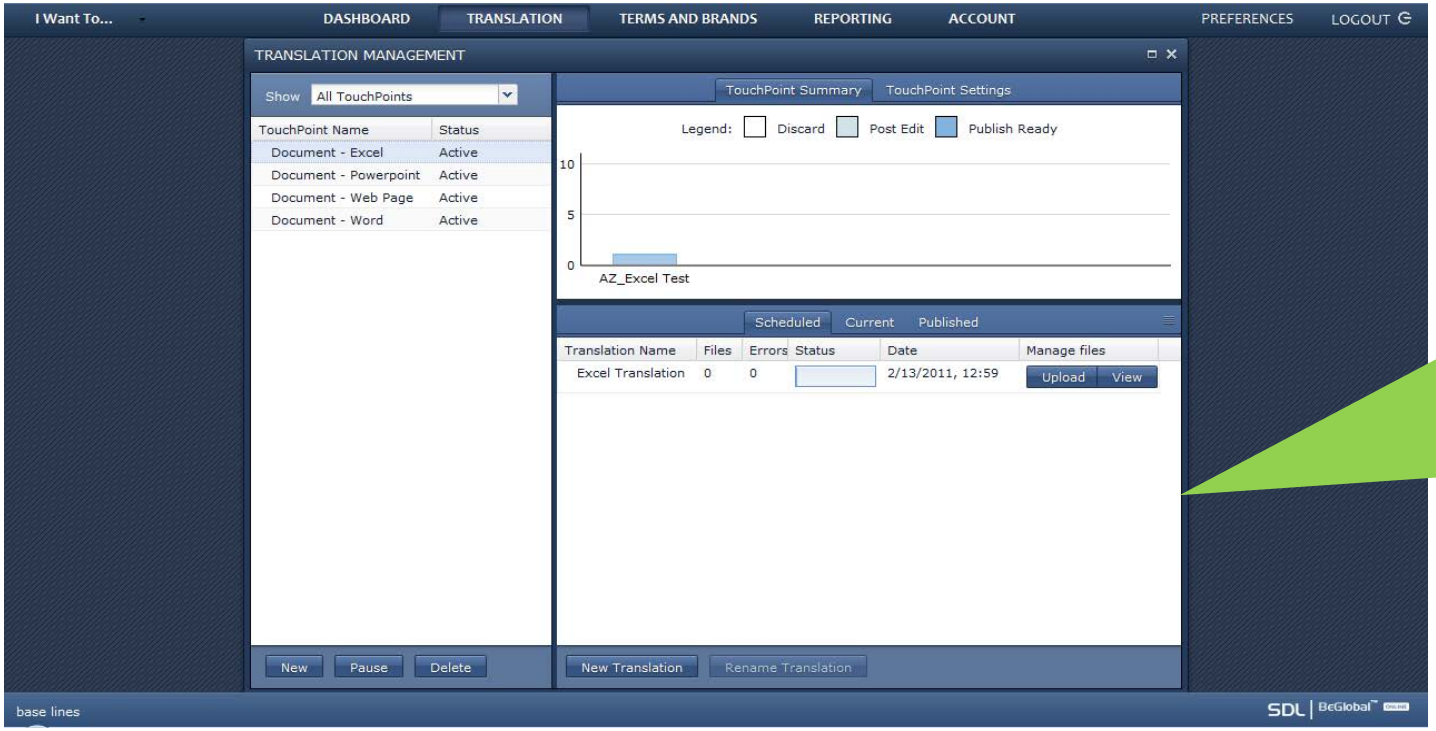
файл будет загружен для перевода

***Шаг 2:***

в диалоге прогресса отразится завершение загрузки

***Шаг 3:***

после загрузки файл будет перемещен в текущую очередь




На экране отразится статус загрузки перевода. По окончании загрузки файл переместится на текущую вкладку "Queue" ("Очередь")

# SDL BeGlobal : Touchpoint (5)

Шаг 1:

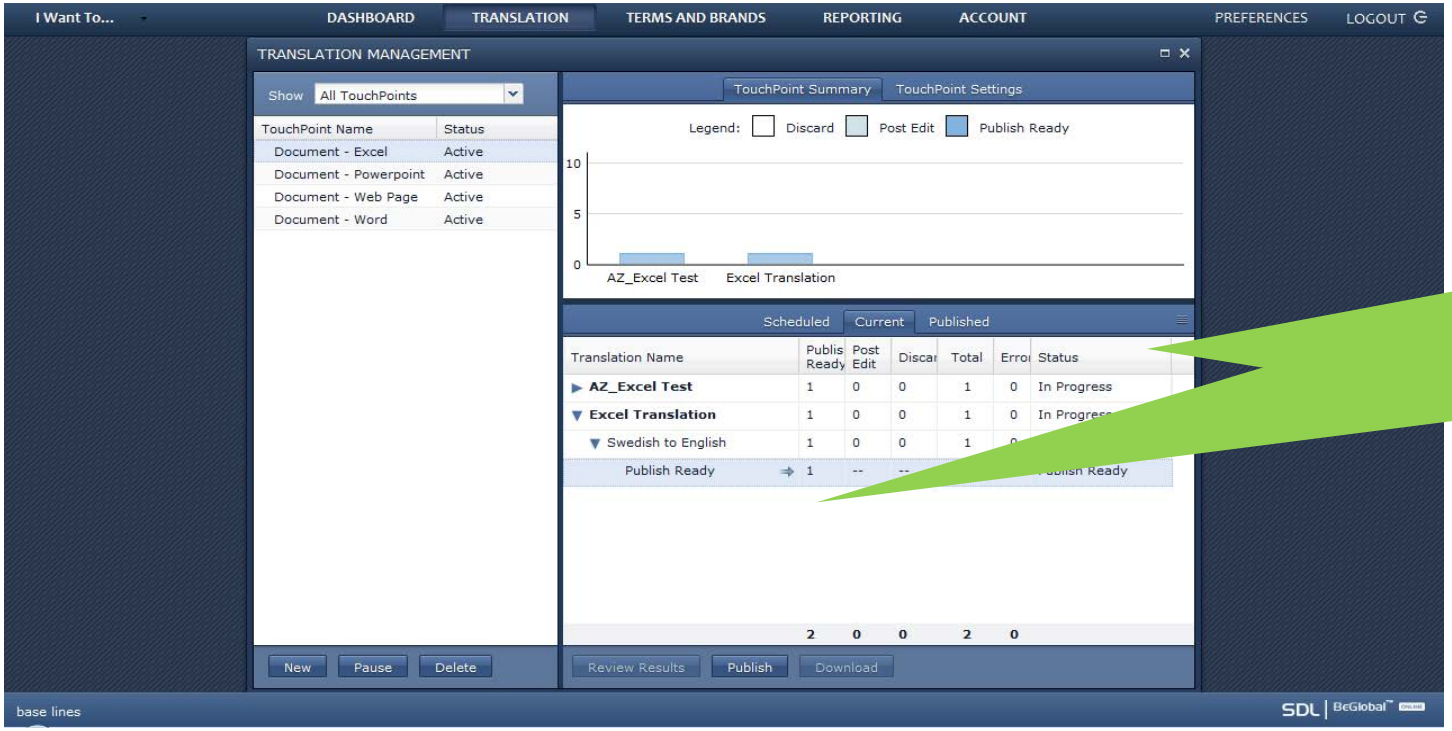
в системе отразятся изменения в ходе работы

***Шаг 2:***

по окончании перевод будет опубликован

***Шаг 3:***

выберите файл, готовый к публикации, и опубликуйте




Изменения вносятся в перевод по ходу работы, по окончании работы документ будет готов к публикации

# SDL BeGlobal : Публикация (1)

**Шаг 1:**

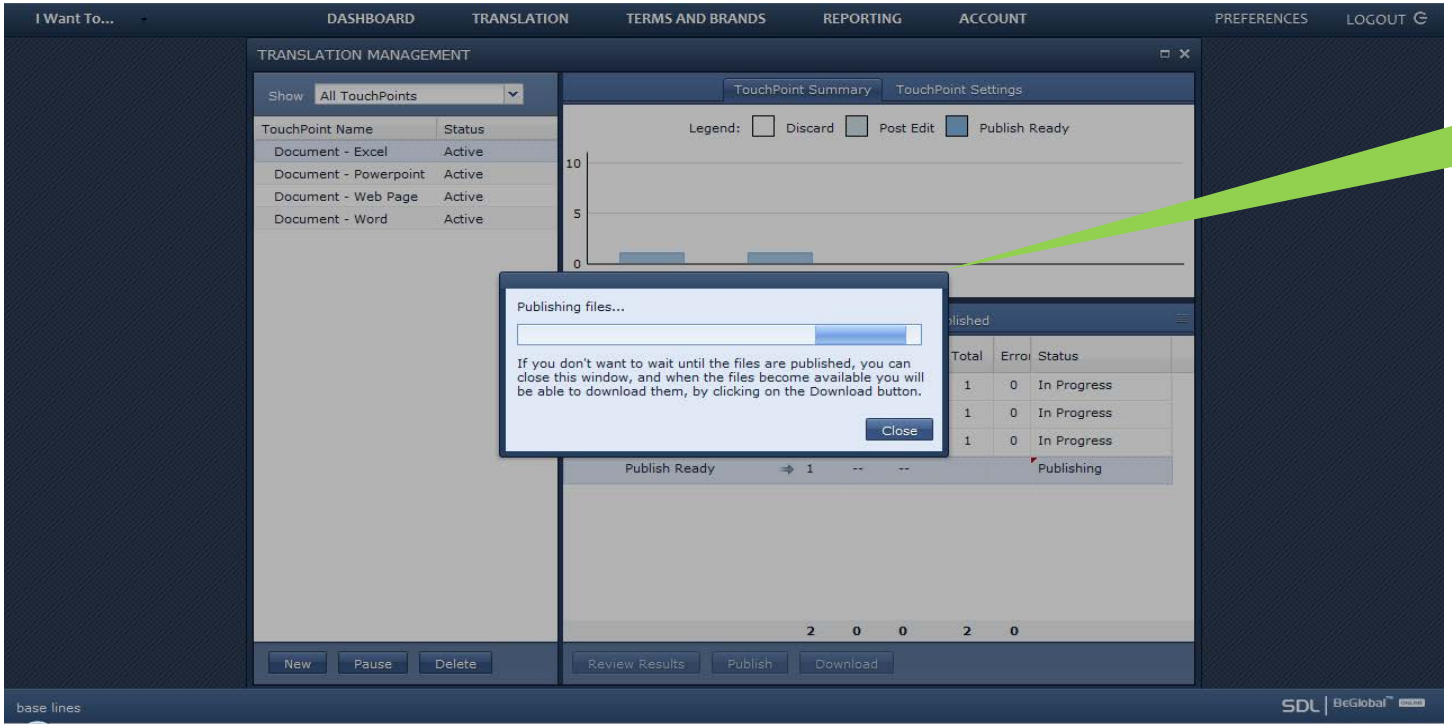
После выбора файла для отправки…

**Шаг 2:**

он будет загружен на локальный диск

**Шаг 3:**

Дождитесь окончания загрузки файла системой




На экране отобразится ход загрузки

# SDL BeGlobal : Публикация (2)

***Шаг 1:***

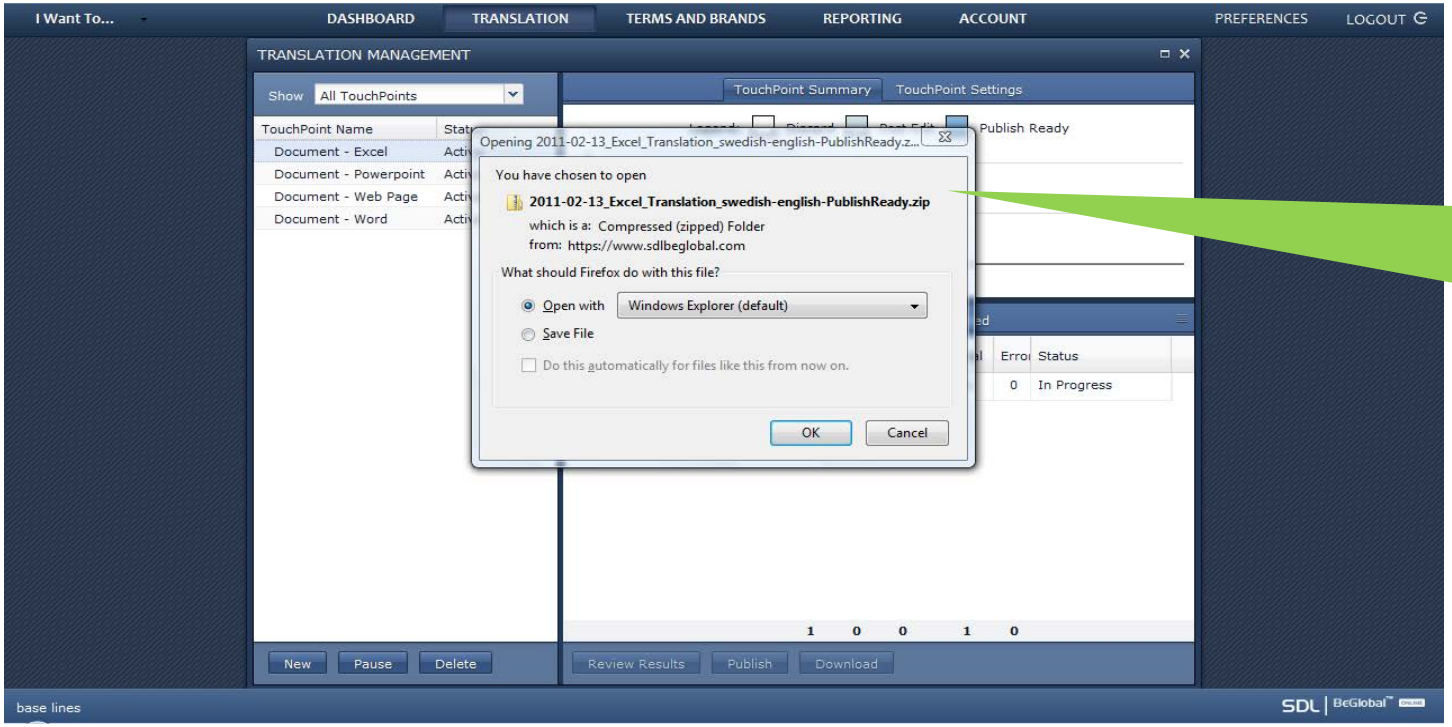
После загрузки можно открыть файл

Шаг 2:

Сохраните файл на локальный диск в формате архива zip

***Шаг 3:***

Архив zip можно сохранить и отправить




Загруженный файл zip можно открыть, сохранить на диск, или отправить

# Конкретный пример: tripadvisor SDL

## Информация о компании

- Ведущее интернет-сообщество путешественников
- Трафик: 25 млн. уникальных посетителей ежемесячно
- Веб-сайты: 10 страны, 8 языков
- Более 20 млн. рецензий и отзывов

## Бизнес-цели внедрения МП

- Увеличение количества публикаций пользователей на английском языке при условии нехватки средств и времени на выполнение перевода человеком
- Высокоскоростной, рентабельный и точный перевод неструктурированных данных в цифровом виде
- Способность управлять миллионами новых статей, созданных пользователями

## Результаты внедрения BeGlobal для 15 языковых пар

- Успешное осуществление перевода, судя по:
  - количеству обращений к сайту;
  - времени пребывания на сайте;
  - количеству просмотров страниц;
  - возросшему числу статей от европейских пользователей благодаря возможности поиска на родном языке
  - увеличению прибыли

# Пользовательский контент: применение

Traduction automatique
Original en anglais

"Outstanding Service & Great Location"

BandBEastMidlands
East Midlands

12 juil. 2008
1/1 ont trouvé cet avis utile

We stayed at the Anastasis appartments for 1 week at the end of June in a Junior Suite. They had been highly recommended by a friend who stayed last year and certainly did not disappoint. Despina the owner & the staff (especially Katerina & Victor) could not be more helpful and were always there with a smile and free Greek... suite

Traduction automatique
Original en anglais

"Fantastic place, perfect holiday"

caroleleeds
london

10 juil. 2008
1/2 ont trouvé cet avis utile

I have to go along with all the other reviews. This place really is fantastic! We have wanted to visit Santorini for a while but never got around to it then on the spur of the moment in April we decided to book Anastasis Apartments. Despina only had one room available and we had to change our dates slightly but... suite

Traduction automatique
Original en anglais

Traduit par : Language Weaver

"Le service est exceptionnel et très bien situé"

BandBEastMidlands
East Midlands

12 juil. 2008
1/1 ont trouvé cet avis utile

Nous avons séjourné au anastasis appartements pendant une semaine à la fin du mois de juin dans une suite junior. Ils avaient été hautement recommandé par un ami qui y étaient allés l'année dernière et ne nous a pas déçus. Despina le propriétaire et le personnel (surtout Katerina et Victor) étaient on ne peut plus serviables et ils étaient toujours... suite

Traduction automatique
Original en anglais

Traduit par : Language Weaver

"Un hôtel fantastique, des vacances parfaites"

caroleleeds
london

10 juil. 2008
1/2 ont trouvé cet avis utile

Je dois aller avec toutes les autres critiques. Cet hôtel est vraiment fantastique ! Nous voulions aller à Santorin depuis un moment mais avons jamais eu le temps d'que la dernière minute en avril Nous avons décidé de réserver anastasis Apartments. Despina n'avait qu'une chambre disponible, nous avons dû changer nos dates légèrement mais quel bon choix ! Nous avions... suite