2012 AMTA 20 Years

The Tenth Biennial Conference of the
Association for Machine Translation in the Americas

Increasing Localization Efficiency with SYSTRAN Hybrid MT Products

John Paul Barraza
Systran

SAN DIEGO, CA
OCTOBER 28– NOVEMBER 1, 2012

This session will cover how to increase localization efficiency with a SYSTRAN desktop product and a server solution. First we will demonstrate how to integrate MT in a localization workflow, interaction with TM matching tools, hands-on MT customization using various tools and dictionaries, and final post-edition using SYSTRAN Premium Translator, a desktop product. We will also walk through the complete cycle of automatic quality improvement using SYSTRAN Training Server, part of the Enterprise Server 7 suite. It covers managing bilingual and monolingual data using Corpus Manager, training hybrid or statistical translation models with Training Manager, and evaluating quality using automatic scoring and side-by-side translation comparison. It also includes other useful tools that automatically extract and validate dictionary entries, and create TMs from unaligned bilingual sentences automatically. Finally, localization efficiency with or without MT integration/customization is compared with the actual cost benefits.

Presenter

- John Paul Barraza, Director of Technical Services at Systran.

SYSTRAN

Increasing Localization Efficiency with SYSTRAN Hybrid MT Products

Presenters

– John Paul Barraza, Director of Services

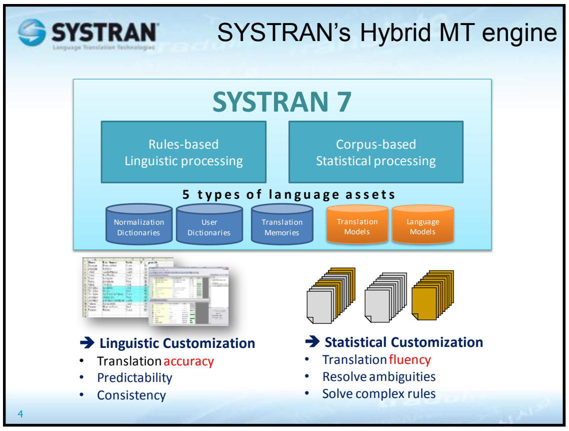
– Philip Staiger, Sr. Technical Trainer

www.systransoft.com

Fast facts

- **Founded in 1968 in CA, now two operating companies**
  - SYSTRAN SA, in Paris, France (parent company)
  - SYSTRAN Software, Inc., in San Diego, CA (original comp.)
  - Publicly traded on NYSE / Euronext (RAN)
- **Business Units**
  - Software Publishing
  - Professional Services
- **Machine Translation (MT) Products and Services**
  - Desktop and Mobile products
  - Server solutions for the enterprise
  - Online services (for portals)
- **Advantages**
  - Scalable, configurable, customizable
  - 43+ years experience

3



SYSTRAN Language Translation Technologies

SYSTRAN's Hybrid MT engine

SYSTRAN 7

Rules-based Linguistic processing

Corpus-based Statistical processing

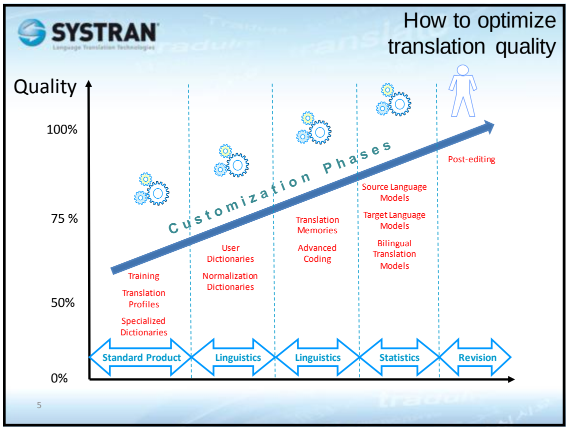
5 types of language assets

Normalization Dictionaries | User Dictionaries | Translation Memories | Translation Models | Language Models

➔ Linguistic Customization
- Translation accuracy
- Predictability
- Consistency

➔ Statistical Customization
- Translation fluency
- Resolve ambiguities
- Solve complex rules

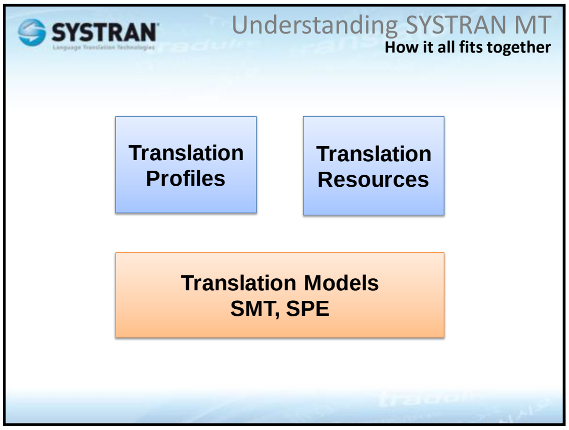
4



SYSTRAN Language Translation Technologies

How to optimize translation quality

5



SYSTRAN Language Translation Technologies

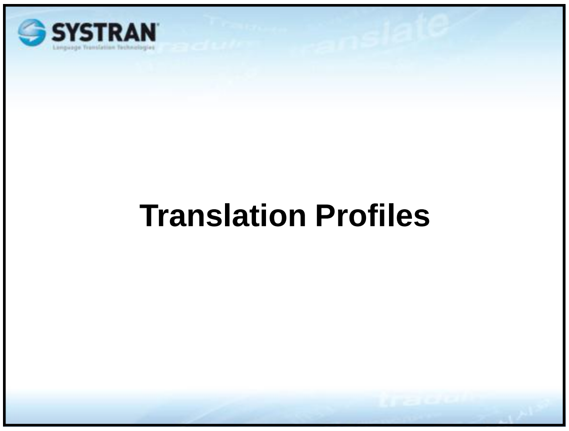
Understanding SYSTRAN MT

How it all fits together

Translation Profiles

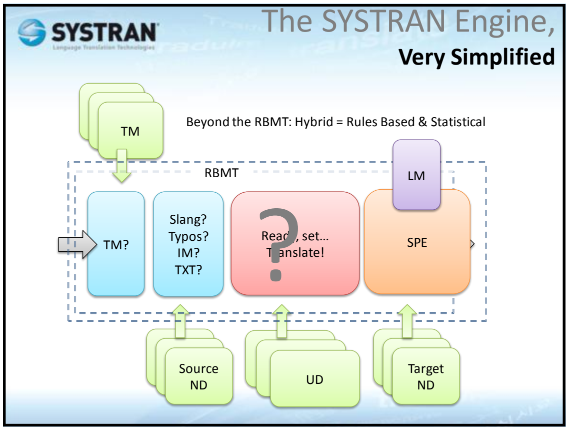
Translation Resources

Translation Models SMT, SPE



SYSTRAN

Translation Profiles



The SYSTRAN Engine, Very Simplified

Beyond the RBMT: Hybrid = Rules Based & Statistical

TM

RBMT

TM?

Slang?
Typos?
IM?
TXT?

?
Ready, set...
Translate!

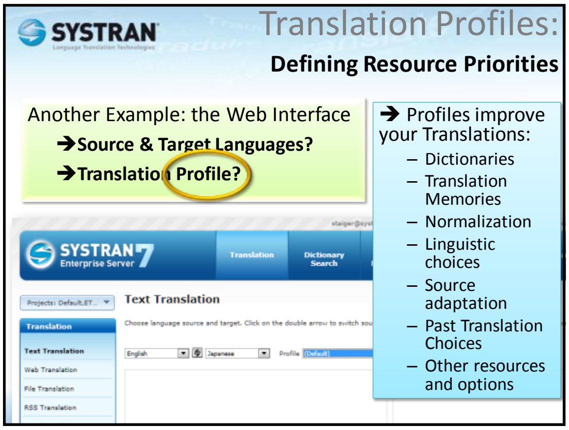
LM

SPE

Source
ND

UD

Target
ND



Translation Profiles: Defining Resource Priorities

Another Example: the Web Interface

→**Source & Target Languages?**

→**Translation Profile?**

staiger@syst

SYSTRAN 7 Enterprise Server

Translation | Dictionary Search

Projects: Default.ET...

Text Translation

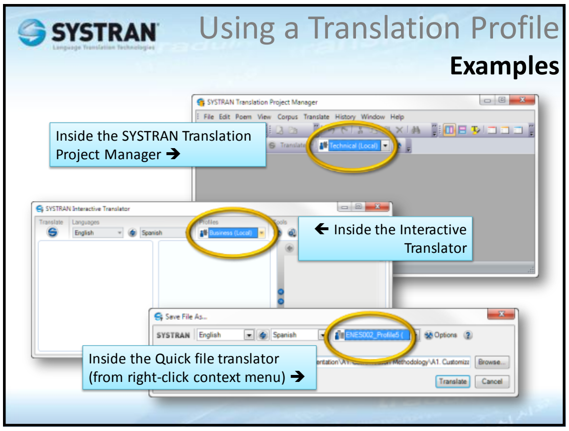
Choose language source and target. Click on the double arrow to switch sou

English | Japanese | Profile (Default)

Translation
- Text Translation
- Web Translation
- File Translation
- RSS Translation

→ Profiles improve your Translations:

- Dictionaries
- Translation Memories
- Normalization
- Linguistic choices
- Source adaptation
- Past Translation Choices
- Other resources and options



Using a Translation Profile

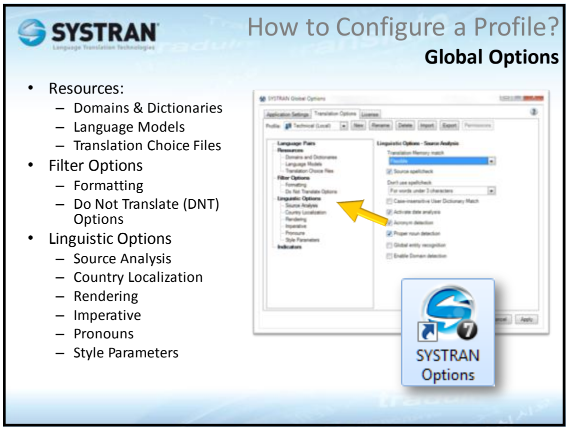
Examples

Inside the SYSTRAN Translation Project Manager ➔

🡐 Inside the Interactive Translator

Inside the Quick file translator (from right-click context menu) ➔

SYSTRAN Translation Project Manager
File Edit Poem View Corpus Translate History Window Help
Technical (Local)
SYSTRAN Interactive Translator
Translate Languages Profiles Tools
English
Spanish
Business (Local)
Save File As...
SYSTRAN English Spanish ENES002_Profile5 (Options
Browse...
Translate Cancel



How to Configure a Profile?

Global Options

- Resources:
  - Domains & Dictionaries
  - Language Models
  - Translation Choice Files
- Filter Options
  - Formatting
  - Do Not Translate (DNT) Options
- Linguistic Options
  - Source Analysis
  - Country Localization
  - Rendering
  - Imperative
  - Pronouns
  - Style Parameters

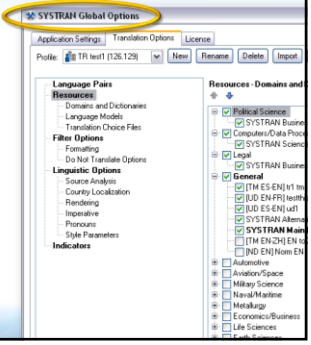
Global Options

Profile Customization

Translation Profiles

What *can* you Customize?

- Resources
  - Dictionaries
- Filter Options
  - Formatting
- Linguistics Options
  - Language specifics
- Indicators
  - Visual cues



Profile: TR test1 (126.129)

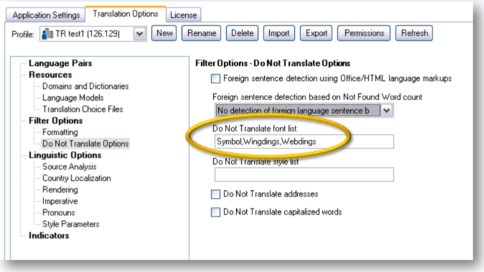
- ☑ Political Science
  - ☑ SYSTRAN Busine
- ☑ Computers/Data Proce
  - ☑ SYSTRAN Scienc
- ☑ Legal
  - ☑ SYSTRAN Busine
- ☑ General
  - ☑ [TM ES-EN] tr1 tm
  - ☑ [UD EN-FR] testth
  - ☑ [UD ES-EN] ud1
  - ☑ SYSTRAN Alterna
  - ☑ **SYSTRAN Main**
  - □ [TM EN-ZH] EN to
  - □ [ND EN] Norm EN
- □ Automotive
- □ Aviation/Space
- □ Military Science
- □ Naval/Maritime
- □ Metallurgy
- □ Economics/Business
- □ Life Sciences
- □ Earth Sciences

Types of Resources

- Dictionaries
  - Several types: UD, TM, ND
  - Within a UD: several types of entries
    - Multilingual entries: Source & Target Language
    - SA : Source Adaptation (grammatical categories)
    - DNT : Do Not Translate (acronyms, proper names…)
- Language Models (TM)
- Translation Choice Files (TRC)

Filter Options

Do Not Translate (DNT)



Profile: TR test1 (126.129)

Filter Options - Do Not Translate Options

□ Foreign sentence detection using Office/HTML language markups

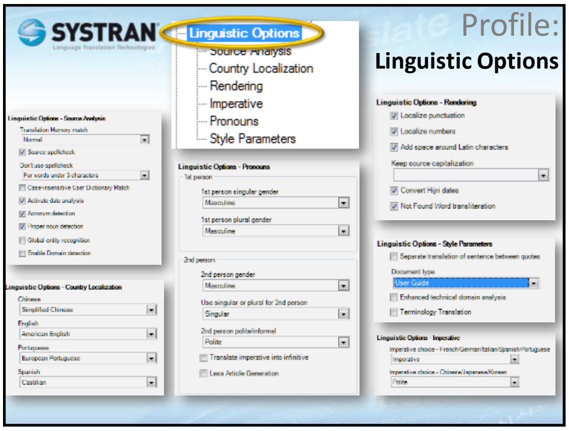
Foreign sentence detection based on Not Found Word count: No detection of foreign language sentence b

Do Not Translate font list: Symbol,Wingdings,Webdings

Do Not Translate style list:

□ Do Not Translate addresses

□ Do Not Translate capitalized words



Profile: Linguistic Options

Linguistic Options
- Source Analysis
- Country Localization
- Rendering
- Imperative
- Pronouns
- Style Parameters

Linguistic Options - Source Analysis

Translation Memory match: Normal

☑ Source spellcheck

Don't use spellcheck: For words under 3 characters

☐ Case-insensitive User Dictionary Match

☑ Activate date analysis

☑ Acronym detection

☑ Proper noun detection

☐ Global entity recognition

☐ Enable Domain detection

Linguistic Options - Country Localization

Chinese: Simplified Chinese

English: American English

Portuguese: European Portuguese

Spanish: Castilian

Linguistic Options - Pronouns

1st person

1st person singular gender: Masculine

1st person plural gender: Masculine

2nd person

2nd person gender: Masculine

Use singular or plural for 2nd person: Singular

2nd person polite/informal: Polite

☐ Translate imperative into infinitive

☐ Less Article Generation

Linguistic Options - Rendering

☑ Localize punctuation

☑ Localize numbers

☑ Add space around Latin characters

Keep source capitalization:

☑ Convert Hijri dates

☑ Not Found Word transliteration

Linguistic Options - Style Parameters

☐ Separate translation of sentence between quotes

Document type: User Guide

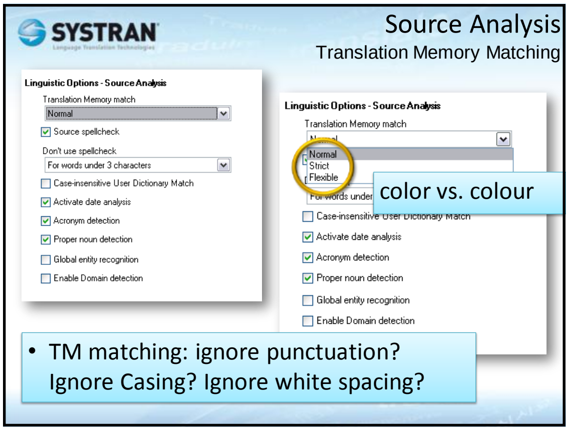
☐ Enhanced technical domain analysis

☐ Terminology Translation

Linguistic Options - Imperative

Imperative choice - French/German/Italian/Spanish/Portuguese: Imperative

Imperative choice - Chinese/Japanese/Korean: Polite



Source Analysis

Translation Memory Matching

Linguistic Options - Source Analysis

Translation Memory match: Normal

☑ Source spellcheck

Don't use spellcheck: For words under 3 characters

☐ Case-insensitive User Dictionary Match

☑ Activate date analysis

☑ Acronym detection

☑ Proper noun detection

☐ Global entity recognition

☐ Enable Domain detection

Linguistic Options - Source Analysis

Translation Memory match

Normal
Strict
Flexible

For words under

color vs. colour

☐ Case-insensitive User Dictionary Match

☑ Activate date analysis

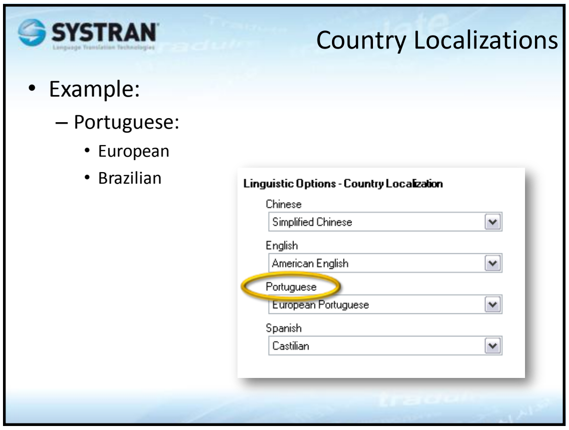
☑ Acronym detection

☑ Proper noun detection

☐ Global entity recognition

☐ Enable Domain detection

- TM matching: ignore punctuation? Ignore Casing? Ignore white spacing?



Country Localizations

- Example:
  - Portuguese:
    - European
    - Brazilian

Linguistic Options - Country Localization

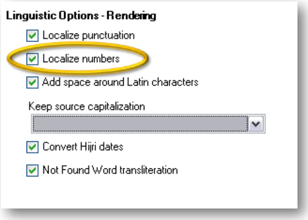
Chinese: Simplified Chinese

English: American English

Portuguese: European Portuguese

Spanish: Castilian

Linguistic Options
Rendering



Linguistic Options - Rendering

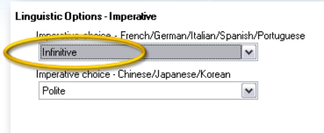
- ☑ Localize punctuation
- ☑ Localize numbers
- ☑ Add space around Latin characters

Keep source capitalization

- ☑ Convert Hijri dates
- ☑ Not Found Word transliteration

- Appearance of translated text
  - Punctuation e.g. 1.00 → 1,00
  - Capitalization

Imperative



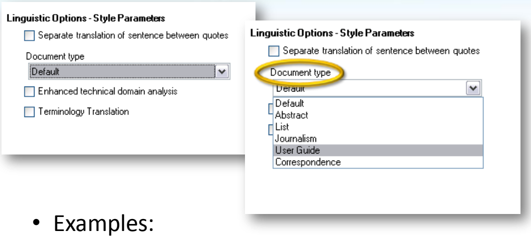
Linguistic Options - Imperative

Imperative choice - French/German/Italian/Spanish/Portuguese: Infinitive

Imperative choice - Chinese/Japanese/Korean: Polite

Example:	Do not lean out!
German infinitive	→ Nicht hinauslehnen!
German imperative	→ Lehnen Sie nicht hinaus!
or informal:	→ Lehne nicht hinaus!
French infinitive:	→ Ne pas se pencher!
Italian (adaptation)	→ E pericoloso sporgersi!

Linguistic Options
Style Parameters



Linguistic Options - Style Parameters

- □ Separate translation of sentence between quotes

Document type: Default

- □ Enhanced technical domain analysis
- □ Terminology Translation

Linguistic Options - Style Parameters

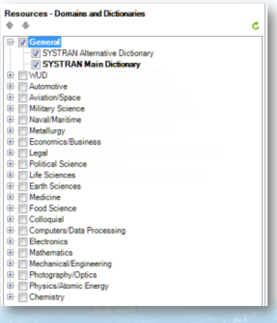
- □ Separate translation of sentence between quotes

Document type: Default

- Examples:
  - Scientific papers vs. News Reports vs. Tech Docs vs. Emails/IMs

Domains:
Standard vs. User-defined

- Always enabled:
  - Main Dictionary, General Domain
- Also included:
  - Alternative Meaning Dictionary
  - 20 special domains, spread over 5 SYSTRAN System dictionaries:
    - Business Dictionary
    - Industries Dictionary
    - Life Sciences Dictionary
    - Science Dictionary
    - Colloquial Dictionary
- Easily Create your own
  - Manually
  - Wizard (WUD)
- Combine & Prioritize



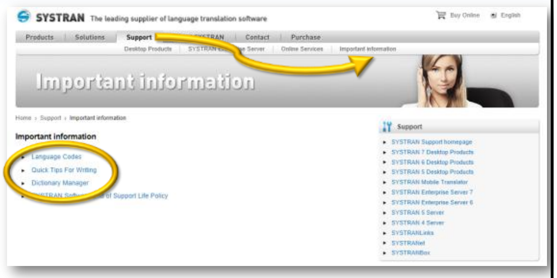
Resources - Domains and Dictionaries

- ☑ General
  - ☑ SYSTRAN Alternative Dictionary
  - ☑ SYSTRAN Main Dictionary
- ☐ WUD
- ☐ Automotive
- ☐ Aviation/Space
- ☐ Military Science
- ☐ Naval/Maritime
- ☐ Metallurgy
- ☐ Economics/Business
- ☐ Legal
- ☐ Political Science
- ☐ Life Sciences
- ☐ Earth Sciences
- ☐ Medicine
- ☐ Food Science
- ☐ Colloquial
- ☐ Computers/Data Processing
- ☐ Electronics
- ☐ Mathematics
- ☐ Mechanical/Engineering
- ☐ Photography/Optics
- ☐ Physics/Atomic Energy
- ☐ Chemistry

Tips for Writing

Source Content Creation
for Better Translation Quality

Tips for Writing
SYSTRAN Web Resources



SYSTRAN The leading supplier of language translation software

Important information

Home › Support › Important information

Important information

- Language Codes
- Quick Tips For Writing
- Dictionary Manager
- SYSTRAN Softw... of Support Life Policy

Support

- SYSTRAN Support homepage
- SYSTRAN 7 Desktop Products
- SYSTRAN 6 Desktop Products
- SYSTRAN 5 Desktop Products
- SYSTRAN Mobile Translator
- SYSTRAN Enterprise Server 7
- SYSTRAN Enterprise Server 6
- SYSTRAN 5 Server
- SYSTRAN 4 Server
- SYSTRANLinks
- SYSTRANet
- SYSTRANBox

SYSTRAN Language Translation Technologies

Tips:
Writing Content for Better Translation Quality

- **Be Direct. Write in a Simple, Clear Manner.**
- **Be Concise and To The Point.**
- **Do Not Leave Out Necessary Words.**
- **Beware of Slang and Colloquialisms.**
- **Insert Proper Punctuation & accents.**
- **Check for Accurate Spelling.**
- **Use Articles Whenever Possible.**
- **Consistent Use of Terminology and Abbreviations.**
- **Maintain a Simple Format.**

SYSTRAN Language Translation Technologies

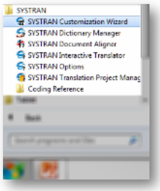
SYSTRAN Customization Wizard

Customizatio n Wizard

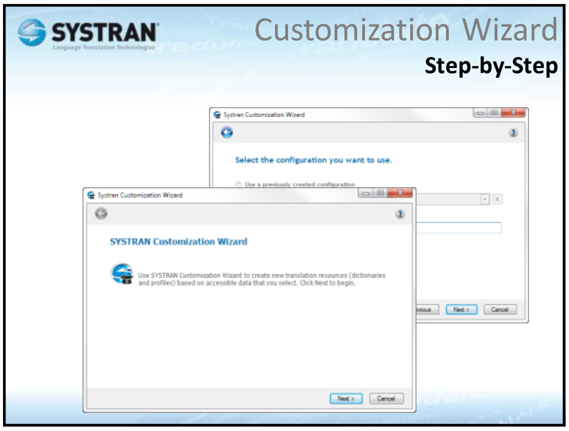
Leveraging your Existing Resources

SYSTRAN Language Translation Technologies

Customization Wizard: **Why?**


SYSTRAN
SYSTRAN Customization Wizard
SYSTRAN Dictionary Manager
SYSTRAN Document Aligner
SYSTRAN Interactive Translator
SYSTRAN Options
SYSTRAN Translation Project Manag
Coding Reference

- Existing Terminology → Dictionaries (WUD)
- Target Language Models



Customization Wizard

Step-by-Step

Systran Customization Wizard

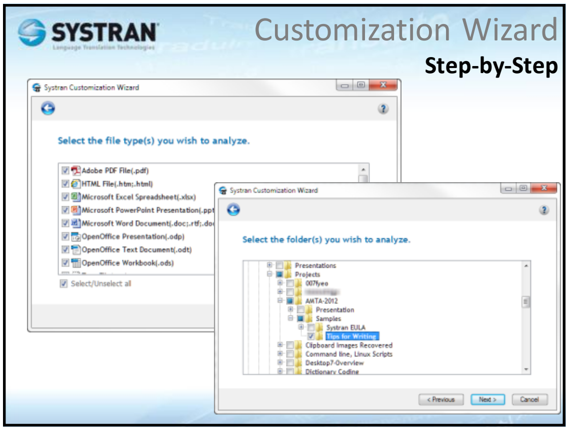
Select the configuration you want to use.

○ Use a previously created configuration

Systran Customization Wizard

SYSTRAN Customization Wizard

Use SYSTRAN Customization Wizard to create new translation resources (dictionaries and profiles) based on accessible data that you select. Click Next to begin.



Customization Wizard

Step-by-Step

Systran Customization Wizard

Select the file type(s) you wish to analyze.

- ☑ Adobe PDF File(.pdf)
- ☑ HTML File(.htm;.html)
- ☑ Microsoft Excel Spreadsheet(.xlsx)
- ☑ Microsoft PowerPoint Presentation(.ppt
- ☑ Microsoft Word Document(.doc;.rtf;.do
- ☑ OpenOffice Presentation(.odp)
- ☑ OpenOffice Text Document(.odt)
- ☑ OpenOffice Workbook(.ods)

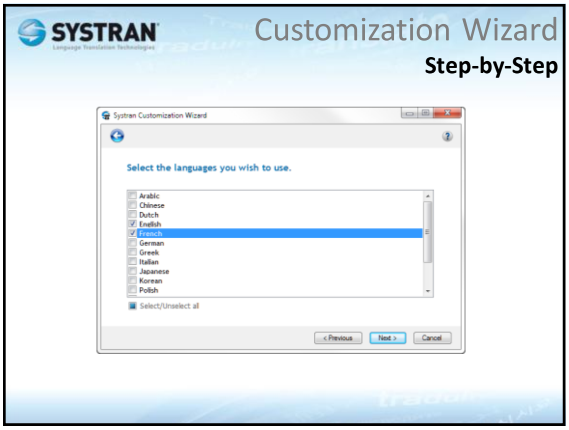
☑ Select/Unselect all

Systran Customization Wizard

Select the folder(s) you wish to analyze.

```
⊞ □ Presentations
⊟ ■ Projects
   ⊞ □ 007fyeo
   ⊞ □ [redacted]
   ⊟ ■ AMTA-2012
      ⊞ □ Presentation
      ⊟ ■ Samples
         ⊞ □ Systran EULA
           ☑ Tips for Writing
   ⊞ □ Clipboard Images Recovered
   ⊞ □ Command line, Linux Scripts
   ⊞ □ Desktop7-Overview
   ⊞ □ Dictionary Coding
```

SYSTRAN
Language Translation Technologies
< Previous
Next >
Cancel



Customization Wizard

Step-by-Step

Systran Customization Wizard

Select the languages you wish to use.

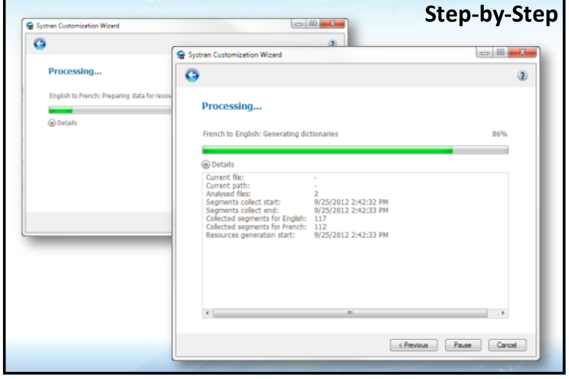
- □ Arabic
- □ Chinese
- □ Dutch
- ☑ English
- ☑ French
- □ German
- □ Greek
- □ Italian
- □ Japanese
- □ Korean
- □ Polish

■ Select/Unselect all

SYSTRAN
Language Translation Technologies
< Previous
Next >
Cancel

Customization Wizard

Step-by-Step



Processing...

English to French: Preparing data for resou

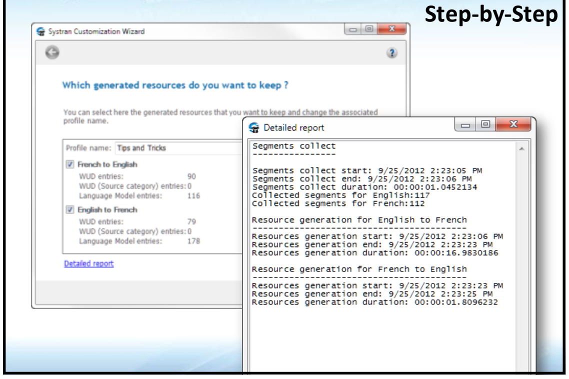
Processing...

French to English: Generating dictionaries 86%

```
Current file:                    -
Current path:                    -
Analysed files:                  2
Segments collect start:          9/25/2012 2:42:32 PM
Segments collect end:            9/25/2012 2:42:33 PM
Collected segments for English:  117
Collected segments for French:   112
Resources generation start:      9/25/2012 2:42:33 PM
```

Customization Wizard

Step-by-Step



Which generated resources do you want to keep ?

You can select here the generated resources that you want to keep and change the associated profile name.

Profile name: Tips and Tricks

☑ French to English

WUD entries:	90
WUD (Source category) entries:	0
Language Model entries:	116

☑ English to French

WUD entries:	79
WUD (Source category) entries:	0
Language Model entries:	178

```
Segments collect
----------------

Segments collect start: 9/25/2012 2:23:05 PM
Segments collect end: 9/25/2012 2:23:06 PM
Segments collect duration: 00:00:01.0452134
Collected segments for English:117
Collected segments for French:112

Resource generation for English to French
-----------------------------------------
Resources generation start: 9/25/2012 2:23:06 PM
Resources generation end: 9/25/2012 2:23:23 PM
Resources generation duration: 00:00:16.9830186

Resource generation for French to English
-----------------------------------------
Resources generation start: 9/25/2012 2:23:23 PM
Resources generation end: 9/25/2012 2:23:25 PM
Resources generation duration: 00:00:01.8096232
```

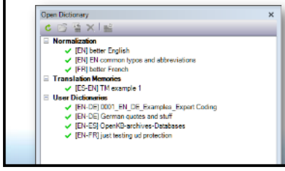
Customization Wizard Output

What are the Resources Created by it?

➔Language Models (non-editable)

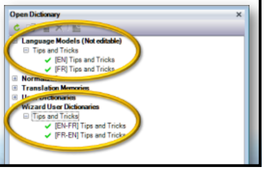
➔Wizard User Dictionaries (editable)

Before

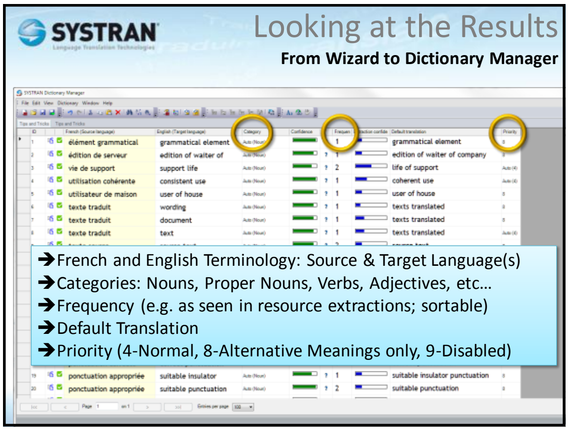


- Normalization
  - ✓ [EN] better English
  - ✓ [EN] EN common typos and abbreviations
  - ✓ [FR] better French
- Translation Memories
  - ✓ [ES-EN] TM example 1
- User Dictionaries
  - ✓ [EN-DE] 0001_EN_DE_Examples_Expert Coding
  - ✓ [EN-DE] German quotes and stuff
  - ✓ [EN-ES] OpenKB-archives-Databases
  - ✓ [EN-FR] just testing ud protection

After



- Language Models (Not editable)
  - Tips and Tricks
    - ✓ [EN] Tips and Tricks
    - ✓ [FR] Tips and Tricks
- Normalization
- Translation Memories
- User Dictionaries
- Wizard User Dictionaries
  - Tips and Tricks
    - ✓ [EN-FR] Tips and Tricks
    - ✓ [FR-EN] Tips and Tricks



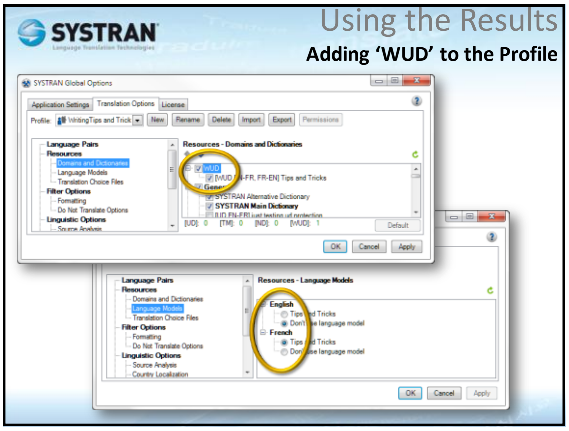
Looking at the Results

From Wizard to Dictionary Manager

SYSTRAN Dictionary Manager

ID	French (Source language)	English (Target language)	Category	Frequen	Default translation	Priority
1	élément grammatical	grammatical element	Auto (Noun)	1	grammatical element	8
2	édition de serveur	edition of waiter of	Auto (Noun)	1	edition of waiter of company	8
3	vie de support	support life	Auto (Noun)	2	life of support	Auto (4)
4	utilisation cohérente	consistent use	Auto (Noun)	1	coherent use	Auto (4)
5	utilisateur de maison	user of house	Auto (Noun)	1	user of house	8
6	texte traduit	wording	Auto (Noun)	1	texts translated	8
7	texte traduit	document	Auto (Noun)	1	texts translated	8
8	texte traduit	text	Auto (Noun)	1	texts translated	Auto (4)
19	ponctuation appropriée	suitable insulator	Auto (Noun)	1	suitable insulator punctuation	8
20	ponctuation appropriée	suitable punctuation	Auto (Noun)	2	suitable punctuation	8

- ➔French and English Terminology: Source & Target Language(s)
- ➔Categories: Nouns, Proper Nouns, Verbs, Adjectives, etc…
- ➔Frequency (e.g. as seen in resource extractions; sortable)
- ➔Default Translation
- ➔Priority (4-Normal, 8-Alternative Meanings only, 9-Disabled)



Using the Results

Adding ‘WUD’ to the Profile

SYSTRAN Global Options

Profile: WritingTips and Trick

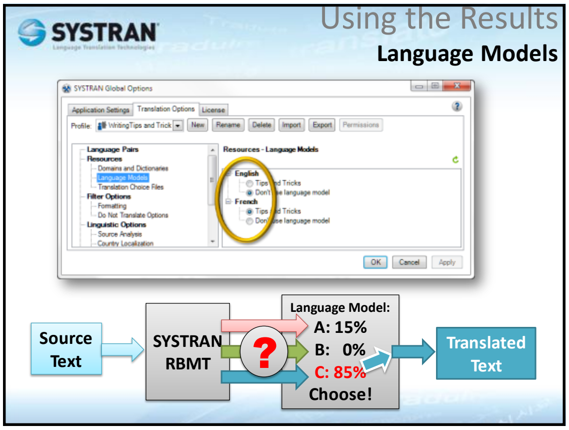
Resources - Domains and Dictionaries

- ☑ WUD
  - ☑ [WUD EN-FR, FR-EN] Tips and Tricks
- ☑ General
  - ☑ SYSTRAN Alternative Dictionary
  - ☑ SYSTRAN Main Dictionary
  - □ [UD EN-FR] just testing ud protection

[UD]: 0 [TM]: 0 [ND]: 0 [WUD]: 1

Resources - Language Models

- English
  - ○ Tips and Tricks
  - ● Don't use language model
- French
  - ● Tips and Tricks
  - ○ Don't use language model



Using the Results

Language Models

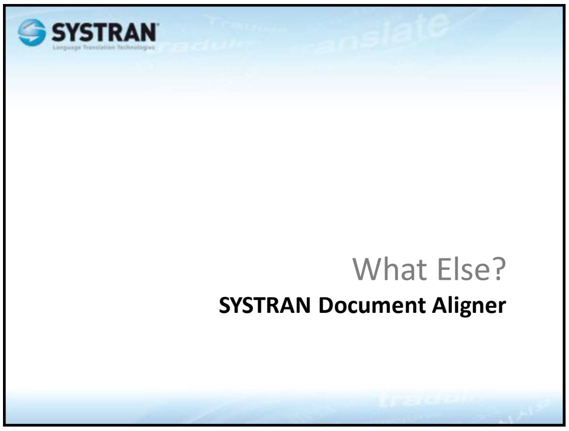
SYSTRAN Global Options

Profile: WritingTips and Trick

Resources - Language Models

- English
  - ○ Tips and Tricks
  - ● Don't use language model
- French
  - ● Tips and Tricks
  - ○ Don't use language model

Language Model:
A: 15%
B: 0%
C: 85%
Choose!

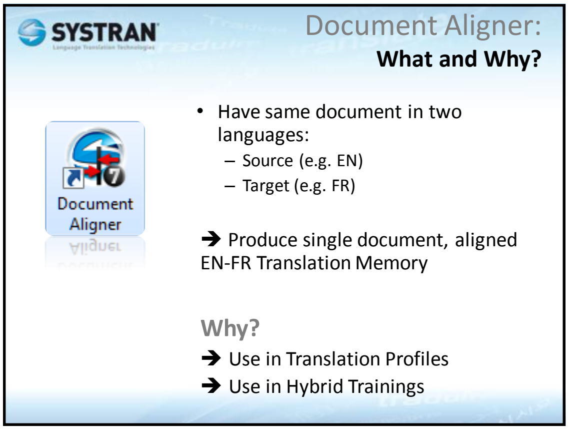


What Else?

SYSTRAN Document Aligner

SYSTRAN
Document Aligner

Source + Target ➔ Bilingual



Document Aligner: **What and Why?**

Document
Aligner

- Have same document in two languages:
  - Source (e.g. EN)
  - Target (e.g. FR)

➔ Produce single document, aligned EN-FR Translation Memory

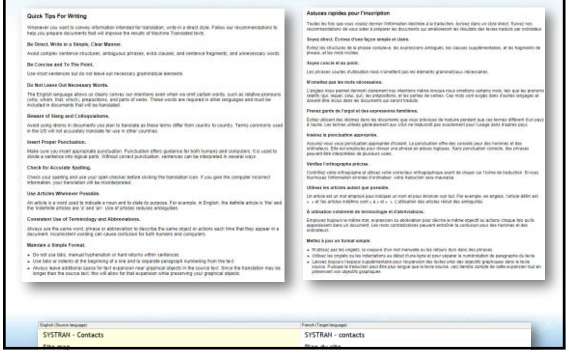
Why?

➔ Use in Translation Profiles

➔ Use in Hybrid Trainings

Document Aligner

An Example



Quick Tips For Writing

Whenever you want to convey information intended for translation, write in a direct style. Follow our recommendations to help you prepare documents that will improve the results of Machine Translated texts.

Be Direct. Write in a Simple, Clear Manner.

Avoid complex sentence structures, ambiguous phrases, extra clauses, and sentence fragments, and unnecessary words.

Be Concise and To The Point.

Use short sentences but do not leave out necessary grammatical elements.

Do Not Leave Out Necessary Words.

The English language allows us clearly convey our intentions even when we omit certain words, such as relative pronouns (who, whom, that, which), prepositions, and parts of verbs. These words are required in other languages and must be included in documents that will be translated.

Beware of Slang and Colloquialisms.

Avoid using idioms in documents you plan to translate as these terms differ from country to country. Terms commonly used in the US will not accurately translate for use in other countries.

Insert Proper Punctuation.

Make sure you insert appropriate punctuation. Punctuation offers guidance for both humans and computers. It is used to divide a sentence into logical parts. Without correct punctuation, sentences can be interpreted in several ways.

Check for Accurate Spelling.

Check your spelling and use your spell-checker before clicking the translation icon. If you give the computer incorrect information, your translation will be misinterpreted.

Use Articles Whenever Possible.

An article is a word used to indicate a noun and to state its purpose. For example, in English, the definite article is 'the' and the 'indefinite articles are 'a' and 'an'. Use of articles reduces ambiguities.

Consistent Use of Terminology and Abbreviations.

Always use the same word, phrase or abbreviation to describe the same object or actions each time that they appear in a document. Inconsistent wording can cause confusion for both humans and computers.

Maintain a Simple Format.

- Do not use tabs, manual hyphenation or hard returns within sentences.
- Use tabs or indents at the beginning of a line and to separate paragraph numbering from the text.
- Always leave additional space for text expansion near graphical objects in the source text. Since the translation may be longer than the source text, this will allow for that expansion while preserving your graphical objects.

Astuces rapides pour l'inscription

Toutes les fois que vous voulez donner l'information destinée à la traduction, écrivez dans un style direct. Suivez nos recommandations de vous aider à préparer les documents qui amélioreront les résultats des textes traduits par ordinateur

Soyez direct. Écrivez d'une façon simple et claire.

Évitez les structures de la phrase complexe, les expressions ambiguës, les clauses supplémentaires, et les fragments de phrase, et les mots inutiles.

Soyez concis et au point.

Les phrases courtes d'utilisation mais n'omettent pas les éléments grammaticaux nécessaires.

N'omettez pas les mots nécessaires.

L'anglais nous permet clairement nos intentions même lorsque nous omettons certains mots, tels que les pronoms relatifs (qui, lequel, celui, qui), les prépositions, et les parties de verbes. Ces mots sont exigés dans d'autres langages et doivent être inclus dans les documents qui seront traduits.

Prenez garde de l'argot et des expressions familières.

Évitez utilisant des idiomes dans les documents que vous prévoyez de traduire pendant que ces termes diffèrent d'un pays à l'autre. Les termes utilisés généralement aux USA ne traduiront pas exactement pour l'usage dans d'autres pays.

Insérez la ponctuation appropriée.

Assurez-vous vous ponctuation appropriée d'insérer. La ponctuation offre des conseils pour des hommes et des ordinateurs. Elle est employée pour diviser une phrase en pièces logiques. Sans ponctuation correcte, des phrases peuvent être interprétées de plusieurs voies.

Vérifiez l'orthographe précise.

Contrôlez votre orthographe et utilisez votre correcteur orthographique avant de cliquer sur l'icône de traduction. Si vous fournissez l'information erronée d'ordinateur, votre traduction sera mauvaise.

Utilisez les articles autant que possible.

Un article est un mot employé pour indiquer un nom et pour énoncer son but. Par exemple, en anglais, l'article défini est « » et les articles indéfinis sont « a » et « ». L'utilisation des articles réduit des ambiguïtés.

À utilisation cohérente de terminologie et d'abréviations.

Employez toujours le même mot, expression ou abréviation pour décrire le même objectif ou actions chaque fois qu'ils apparaissent dans un document. Les mots contradictoires peuvent entraîner la confusion pour des hommes et des ordinateurs.

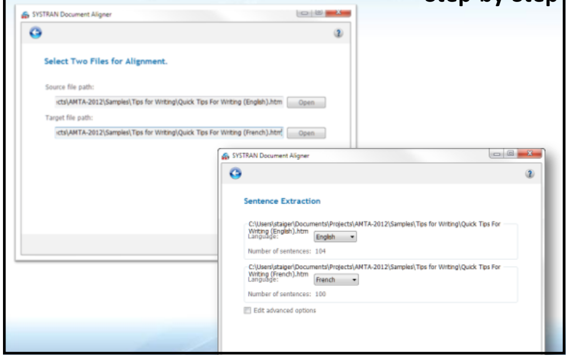
Mettez à jour un format simple.

- N'utilisez pas les onglets, la coupure d'un mot manuelle ou les retours durs dans des phrases.
- Utilisez les onglets ou les indentations au début d'une ligne et pour séparer la numérotation de paragraphe du texte.
- Laissez toujours l'espace supplémentaire pour l'expansion des textes près des objectifs graphiques dans le texte source. Puisque la traduction peut être plus longue que le texte source, ceci tiendra compte de cette expansion tout en préservant vos objectifs graphiques.

English (Source language)	French (Target language)
SYSTRAN - Contacts	SYSTRAN - contacts
Site map	Plan du site

Document Aligner

Step-by-Step



SYSTRAN Document Aligner

Select Two Files for Alignment.

Source file path: cts\AMTA-2012\Samples\Tips for Writing\Quick Tips For Writing (English).htm

Target file path: cts\AMTA-2012\Samples\Tips for Writing\Quick Tips For Writing (French).htm

SYSTRAN Document Aligner

Sentence Extraction

C:\Users\staiger\Documents\Projects\AMTA-2012\Samples\Tips for Writing\Quick Tips For Writing (English).htm

Language: English

Number of sentences: 104

C:\Users\staiger\Documents\Projects\AMTA-2012\Samples\Tips for Writing\Quick Tips For Writing (French).htm

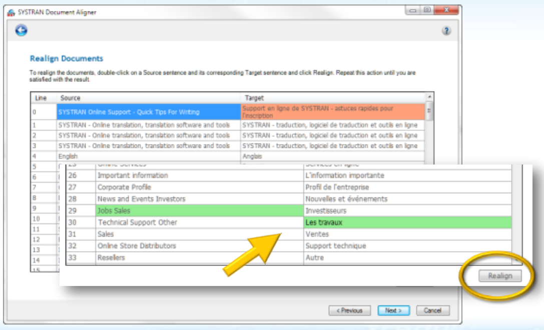
Language: French

Number of sentences: 100

☐ Edit advanced options

Document Aligner

Step-by-Step



SYSTRAN Document Aligner

Realign Documents

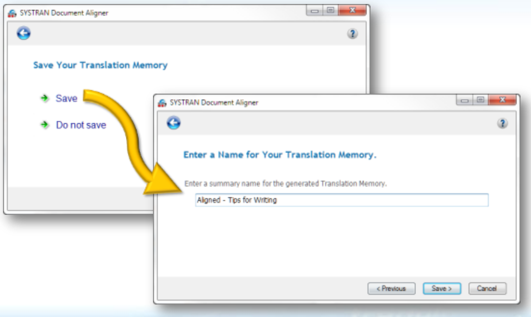
To realign the documents, double-click on a Source sentence and its corresponding Target sentence and click Realign. Repeat this action until you are satisfied with the result.

Line	Source	Target
0	SYSTRAN Online Support - Quick Tips For Writing	Support en ligne de SYSTRAN - astuces rapides pour l'inscription
1	SYSTRAN - Online translation, translation software and tools	SYSTRAN - traduction, logiciel de traduction et outils en ligne
2	SYSTRAN - Online translation, translation software and tools	SYSTRAN - traduction, logiciel de traduction et outils en ligne
3	SYSTRAN - Online translation, translation software and tools	SYSTRAN - traduction, logiciel de traduction et outils en ligne
4	English	Anglais

26	Important information	L'information importante
27	Corporate Profile	Profil de l'entreprise
28	News and Events Investors	Nouvelles et événements
29	Jobs Sales	Investisseurs
30	Technical Support Other	Les travaux
31	Sales	Ventes
32	Online Store Distributors	Support technique
33	Resellers	Autre

Document Aligner

Step-by-Step



SYSTRAN Document Aligner

Save Your Translation Memory

- Save
- Do not save

SYSTRAN Document Aligner

Enter a Name for Your Translation Memory.

Enter a summary name for the generated Translation Memory.

Aligned - Tips for Writing

What's Next?

Working with Dictionaries

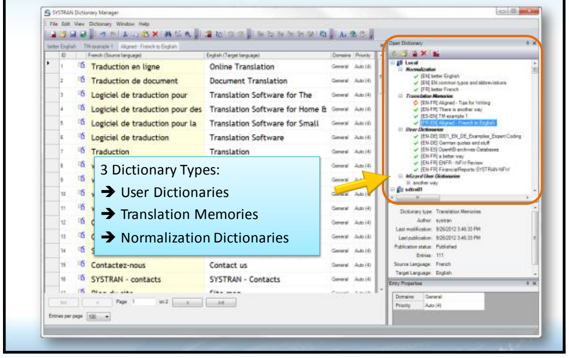
SYSTRAN Dictionary Manager

SYSTRAN Dictionary Manager

Terminology Resource and Translation Memory Management with SDM

SYSTRAN Dictionary Manager

SDM Overview



SYSTRAN Dictionary Manager
File Edit View Dictionary Window Help
better English
TM example 1
Aligned - French to English

ID	French (Source language)	English (Target language)	Domains	Priority
1	Traduction en ligne	Online Translation	General	Auto (4)
2	Traduction de document	Document Translation	General	Auto (4)
3	Logiciel de traduction pour	Translation Software for The	General	Auto (4)
4	Logiciel de traduction pour des	Translation Software for Home &	General	Auto (4)
5	Logiciel de traduction pour la	Translation Software for Small	General	Auto (4)
6	Logiciel de traduction	Translation Software	General	Auto (4)
7	Traduction	Translation	General	Auto (4)
15	Contactez-nous	Contact us	General	Auto (4)
16	SYSTRAN - contacts	SYSTRAN - Contacts	General	Auto (4)

Open Dictionary

- Local
  - Normalization
    - [EN] better English
    - [EN] EN common typos and abbreviations
    - [FR] better French
  - Translation Memories
    - [EN-FR] Aligned - Tips for Writing
    - [EN-FR] There is another way
    - [ES-EN] TM example 1
    - [FR-EN] Aligned - French to English
  - User Dictionaries
    - [EN-DE] 0001_EN_DE_Examples_Expert Coding
    - [EN-DE] German quotes and stuff
    - [EN-ES] OpenKB-archives-Databases
    - [EN-FR] a better way
    - [EN-FR] ENFR - NFrr Review
    - [EN-FR] FinancialReports-SYSTRAN-NFrr
  - Wizard User Dictionaries
    - another way
- sdtrs01

Dictionary type: Translation Memories
Author: systran
Last modification: 9/26/2012 3:46:33 PM
Last publication: 9/26/2012 3:46:33 PM
Publication status: Published
Entries: 111
Source Language: French
Target Language: English

Entry Properties

Domains: General
Priority: Auto (4)

Page 1 on 2
Entries per page: 100

3 Dictionary Types:
➔ User Dictionaries
➔ Translation Memories
➔ Normalization Dictionaries

User Dictionaries

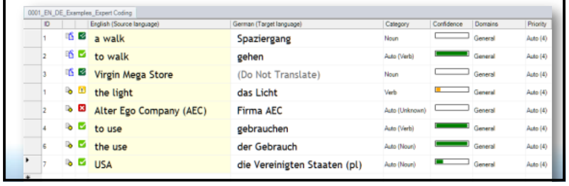
Terminology

➔ Proper Nouns, Acronyms, Enterprise Vocabulary

- Bi-lingual, multi-lingual
- Grammatical Category
- Priority
- Domain(s)
- Sort it by Confidence
- Status: ? ✓ ✓ ⚠ ✗

➔ Example: [EN-FR]

to use	utiliser
use	usage (m)



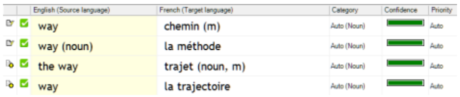
0001_EN_DE_Examples_Expert Coding

ID		English (Source language)	German (Target language)	Category	Confidence	Domains	Priority
1	✓	a walk	Spaziergang	Noun		General	Auto (4)
2	✓	to walk	gehen	Auto (Verb)		General	Auto (4)
3	✓	Virgin Mega Store	(Do Not Translate)	Noun		General	Auto (4)
1	⚠	the light	das Licht	Verb		General	Auto (4)
2	✗	Alter Ego Company (AEC)	Firma AEC	Auto (Unknown)		General	Auto (4)
4	✓	to use	gebrauchen	Auto (Verb)		General	Auto (4)
6	✓	the use	der Gebrauch	Auto (Noun)		General	Auto (4)
7	✓	USA	die Vereinigten Staaten (pl)	Auto (Noun)		General	Auto (4)

User Dictionaries

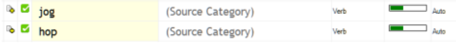
3 Entry Types

➔Multilingual



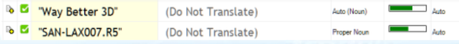
	English (Source language)	French (Target language)	Category	Confidence	Priority
✓	way	chemin (m)	Auto (Noun)		Auto
✓	way (noun)	la méthode	Auto (Noun)		Auto
✓	the way	trajet (noun, m)	Auto (Noun)		Auto
✓	way	la trajectoire	Auto (Noun)		Auto

➔ Source Category



✓	jog	(Source Category)	Verb		Auto
✓	hop	(Source Category)	Verb		Auto

➔ Do Not Translate

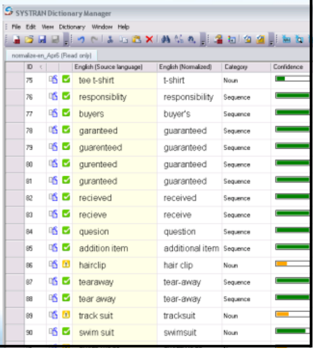


✓	"Way Better 3D"	(Do Not Translate)	Auto (Noun)		Auto
✓	"SAN-LAX007.R5"	(Do Not Translate)	Proper Noun		Auto

Normalization Dictionaries

Synonyms (source or target language are the same)

- Initials, Acronyms and Colloquial (chat)
- Common typos

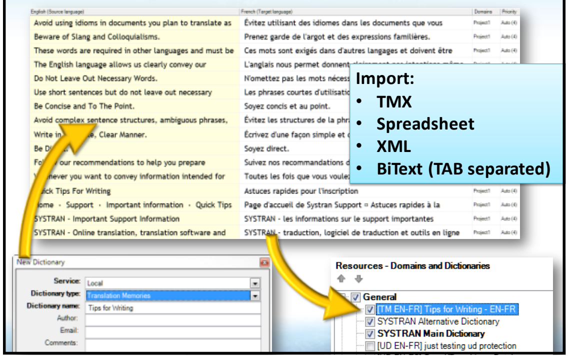


SYSTRAN Dictionary Manager

normalize-en_Apr6 (Read only)

ID	English (Source language)	English (Normalized)	Category
75	tee t-shirt	t-shirt	Noun
76	responsiblity	responsibility	Sequence
77	buyers	buyer's	Sequence
78	garanteed	guaranteed	Sequence
79	guarenteed	guaranteed	Sequence
80	gurenteed	guaranteed	Sequence
81	guranteed	guaranteed	Sequence
82	recieved	received	Sequence
83	recieve	receive	Sequence
84	quesion	question	Sequence
85	addition item	additional item	Sequence
86	hairclip	hair clip	Noun
87	tearaway	tear-away	Sequence
88	tear away	tear-away	Sequence
89	track suit	tracksuit	Noun
90	swim suit	swimsuit	Noun

Translation Memories



English (Source language)	French (Target language)	Domains	Priority
Avoid using idioms in documents you plan to translate as	Évitez utilisant des idiomes dans les documents que vous	Project1	Auto (4)
Beware of Slang and Colloquialisms.	Prenez garde de l'argot et des expressions familières.	Project1	Auto (4)
These words are required in other languages and must be	Ces mots sont exigés dans d'autres langages et doivent être	Project1	Auto (4)
The English language allows us clearly convey our	L'anglais nous permet donnent		
Do Not Leave Out Necessary Words.	N'omettez pas les mots nécess		
Use short sentences but do not leave out necessary	Les phrases courtes d'utilisatio		
Be Concise and To The Point.	Soyez concis et au point.		
Avoid complex sentence structures, ambiguous phrases,	Évitez les structures de la phr		
Write in e, Clear Manner.	Écrivez d'une façon simple et		
Be D	Soyez direct.		
Fo our recommendations to help you prepare	Suivez nos recommandations d		
never you want to convey information intended for	Toutes les fois que vous voulez		
ck Tips For Writing	Astuces rapides pour l'inscription	Project1	Auto (4)
ome · Support · Important information · Quick Tips	Page d'accueil de Systran Support ¤ Astuces rapides à la	Project1	Auto (4)
SYSTRAN - Important Support Information	SYSTRAN - les informations sur le support importantes	Project1	Auto (4)
SYSTRAN - Online translation, translation software and	SYSTRAN - traduction, logiciel de traduction et outils en ligne	Project1	Auto (4)

New Dictionary

Service: Local

Dictionary type: Translation Memories

Dictionary name: Tips for Writing

Author:

Email:

Comments:

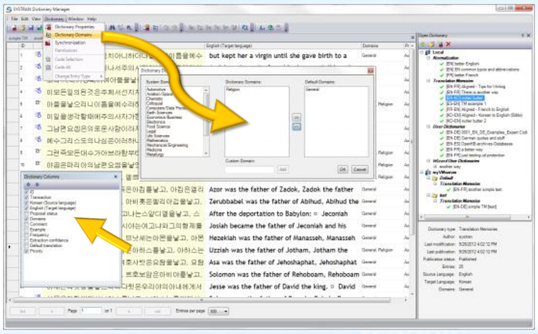
Resources - Domains and Dictionaries

- ☑ General
  - ☑ [TM EN-FR] Tips for Writing - EN-FR
  - ☑ SYSTRAN Alternative Dictionary
  - ☑ SYSTRAN Main Dictionary
  - □ [UD EN-FR] just testing ud protection

Import:

- **TMX**
- **Spreadsheet**
- **XML**
- **BiText (TAB separated)**

Domain Management



SYSTRAN Dictionary Manager

but kept her a virgin until she gave birth to a

Azor was the father of Zadok, Zadok the father

Zerubbabel was the father of Abihud, Abihud the

After the deportation to Babylon: ¤ Jeconiah

Josiah became the father of Jeconiah and his

Hezekiah was the father of Manasseh, Manasseh

Uzziah was the father of Jotham, Jotham the

Asa was the father of Jehoshaphat, Jehoshaphat

Solomon was the father of Rehoboam, Rehoboam

Jesse was the father of David the king. ¤ David

Dictionary type: Translation Memories

Author: systran

Last modification: 9/26/2012 4:02:12 PM

Last publication: 9/26/2012 4:02:12 PM

Publication status: Published

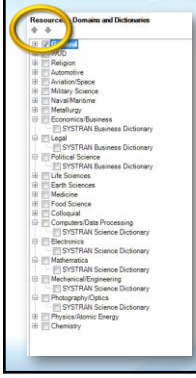
Entries: 25

Source Language: English

Target Language: Korean

Domains: General

Domain Management



☑ General
□ WUD
□ Religion
□ Automotive
□ Aviation/Space
□ Military Science
□ Naval/Maritime
□ Metallurgy
□ Economics/Business
  □ SYSTRAN Business Dictionary
□ Legal
  □ SYSTRAN Business Dictionary
□ Political Science
  □ SYSTRAN Business Dictionary
□ Life Sciences
□ Earth Sciences
□ Medicine
□ Food Science
□ Colloquial
□ Computers/Data Processing
  □ SYSTRAN Science Dictionary
□ Electronics
  □ SYSTRAN Science Dictionary
□ Mathematics
  □ SYSTRAN Science Dictionary
□ Mechanical/Engineering
  □ SYSTRAN Science Dictionary
□ Photography/Optics
  □ SYSTRAN Science Dictionary
□ Physics/Atomic Energy
□ Chemistry

Examples:

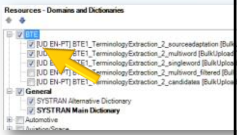
- SYSTRAN Business Dictionary:
  - Economics/Business
  - Legal
  - Political Science
- SYSTRAN Science Dictionary:
  - Computers/Data Processing
  - Electronics
  - Mathematics
  - Mechanical/Engineering
  - Photography/Optics
  - Physics/Atomic Energy
  - Chemistry

➔ Does staking order matter?

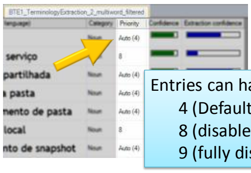
Profiles & Dictionaries

Stacking & Priority Rules

Profile Stacking Order:
➔Top Dictionaries seen first



☑ BTE
  ☑ [UD EN-PT] BTE1_TerminologyExtraction_2_sourceadaptation [Bulk
  ☑ [UD EN-PT] BTE1_TerminologyExtraction_2_multiword [BulkUpload
  ☑ [UD EN-PT] BTE1_TerminologyExtraction_2_singleword [BulkUploa
  □ [UD EN-PT] BTE1_TerminologyExtraction_2_multiword_filtered [Bul
  □ [UD EN-PT] BTE1_TerminologyExtraction_2_candidates [BulkUploa
☑ General
  ☑ SYSTRAN Alternative Dictionary
  ☑ SYSTRAN Main Dictionary
□ Automotive
□ Aviation/Space


BTE1_TerminologyExtraction_2_multiword_filtered

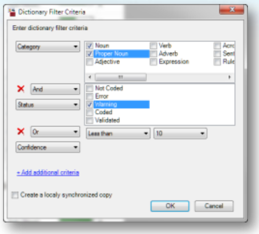
language)	Category	Priority	Confidence	Extraction confidence
	Noun	Auto (4)		
serviço		8		
partilhada	Noun	Auto (4)		
a pasta	Noun	Auto (4)		
mento de pasta	Noun	Auto (4)		
local	Noun	8		
nto de snapshot	Noun	Auto (4)		

Entries can have their own priority!
4 (Default)
8 (disabled & visible to Alternative Meaning)
9 (fully disabled & invisible)

Resource Extraction, Dictionary Validation?
➔Changing entry priority to enable/disable

SDM

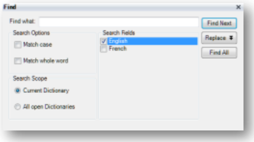
Advanced Search Tools


Dictionary Filter Criteria
Enter dictionary filter criteria
+ Add additional criteria
OK
Cancel

Category: ☑ Noun ☑ Proper Noun □ Adjective □ Verb □ Adverb □ Expression □ Acro □ Sent □ Rule
And
Status: □ Not Coded □ Error ☑ Warning □ Coded □ Validated
Or
Confidence: Less than 10
□ Create a localy synchronized copy

Dictionary Filtering

➔ Entries with Problems?

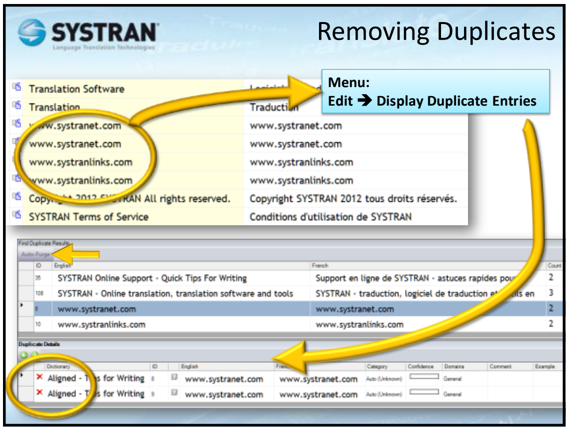


Find what:
Search Options: □ Match case □ Match whole word
Search Fields: ☑ English □ French
Search Scope: ● Current Dictionary ○ All open Dictionaries

Find

➔Entries with Space(s)

➔Separate multi-word entries from single-word entries



Removing Duplicates

Menu:
Edit ➔ Display Duplicate Entries

Translation Software	Logiciel de traduction
Translation	Traduction
www.systranet.com	www.systranet.com
www.systranet.com	www.systranet.com
www.systranlinks.com	www.systranlinks.com
www.systranlinks.com	www.systranlinks.com
Copyright 2012 SYSTRAN All rights reserved.	Copyright SYSTRAN 2012 tous droits réservés.
SYSTRAN Terms of Service	Conditions d'utilisation de SYSTRAN

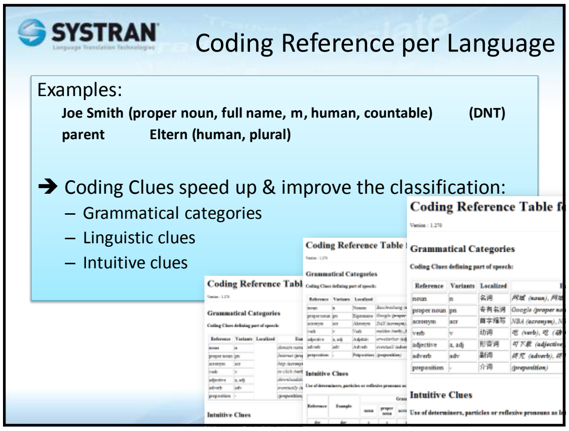
Find Duplicate Results

Auto-Purge

ID	English	French	Count
35	SYSTRAN Online Support - Quick Tips For Writing	Support en ligne de SYSTRAN - astuces rapides pou	2
108	SYSTRAN - Online translation, translation software and tools	SYSTRAN - traduction, logiciel de traduction et outils en	3
8	www.systranet.com	www.systranet.com	2
10	www.systranlinks.com	www.systranlinks.com	2

Duplicate Details

	Dictionary	ID		English	French	Category	Confidence	Domains	Comment	Example
✕	Aligned - Tips for Writing	8	□	www.systranet.com	www.systranet.com	Auto (Unknown)		General		
✕	Aligned - Tips for Writing	9	□	www.systranet.com	www.systranet.com	Auto (Unknown)		General		



Coding Reference per Language

Examples:

Joe Smith (proper noun, full name, m, human, countable) (DNT)

parent Eltern (human, plural)

➔ Coding Clues speed up & improve the classification:

- Grammatical categories
- Linguistic clues
- Intuitive clues

Coding Reference Table f

Version : 1.270

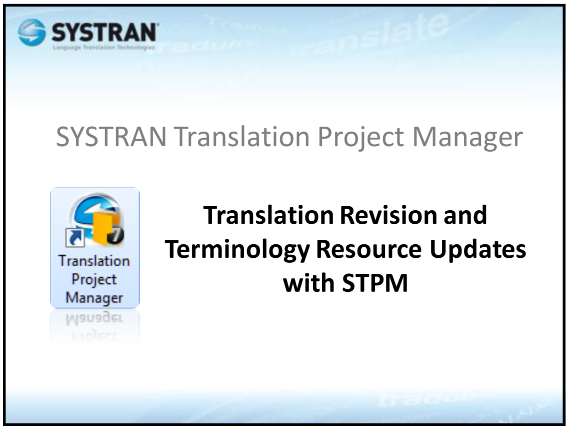
Grammatical Categories

Coding Clues defining part of speech:

Reference	Variants	Localized
noun	n	名词
proper noun	pn	专有名词
acronym	acr	首字缩写
verb	v	动词
adjective	a, adj	形容词
adverb	adv	副词
preposition	-	介词

Intuitive Clues

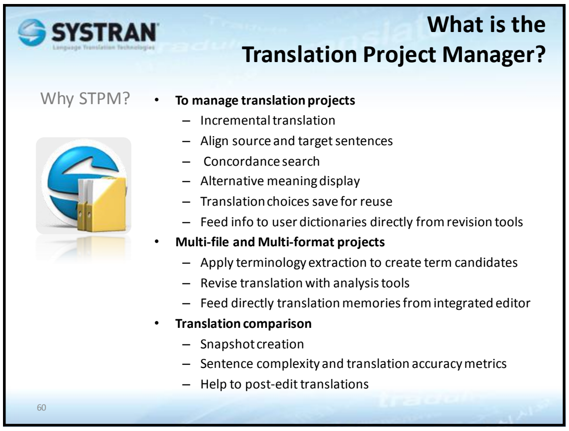
Use of determiners, particles or reflexive pronouns as



SYSTRAN Translation Project Manager

Translation Project Manager

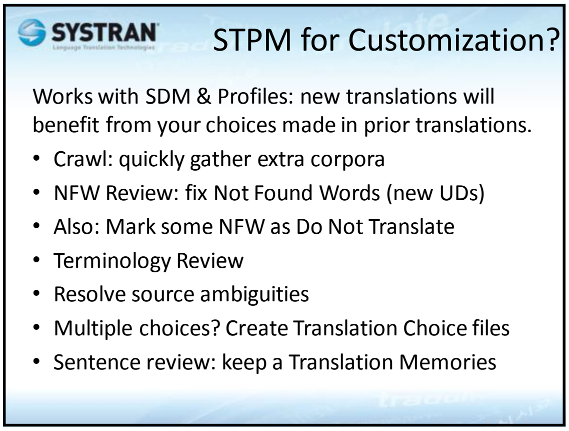
Translation Revision and Terminology Resource Updates with STPM



What is the Translation Project Manager?

Why STPM?

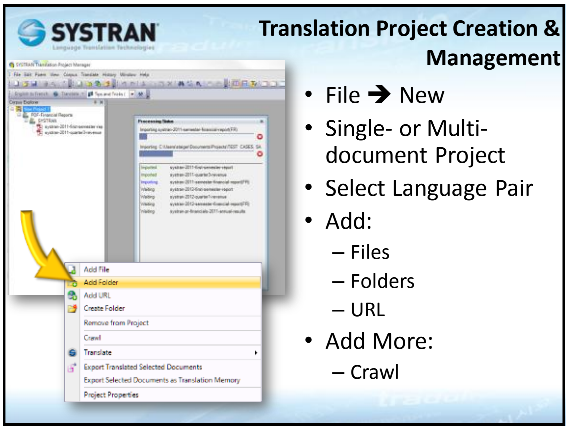
- **To manage translation projects**
  - Incremental translation
  - Align source and target sentences
  - Concordance search
  - Alternative meaning display
  - Translation choices save for reuse
  - Feed info to user dictionaries directly from revision tools
- **Multi-file and Multi-format projects**
  - Apply terminology extraction to create term candidates
  - Revise translation with analysis tools
  - Feed directly translation memories from integrated editor
- **Translation comparison**
  - Snapshot creation
  - Sentence complexity and translation accuracy metrics
  - Help to post-edit translations



STPM for Customization?

Works with SDM & Profiles: new translations will benefit from your choices made in prior translations.

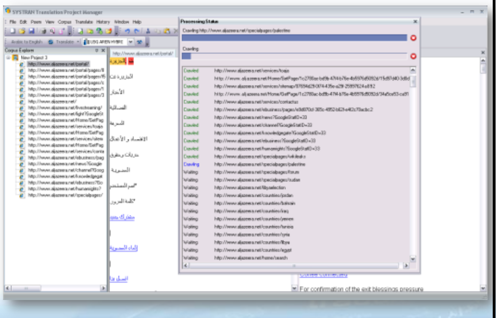
- Crawl: quickly gather extra corpora
- NFW Review: fix Not Found Words (new UDs)
- Also: Mark some NFW as Do Not Translate
- Terminology Review
- Resolve source ambiguities
- Multiple choices? Create Translation Choice files
- Sentence review: keep a Translation Memories



Translation Project Creation & Management

- File ➔ New
- Single- or Multi-document Project
- Select Language Pair
- Add:
  - Files
  - Folders
  - URL
- Add More:
  - Crawl

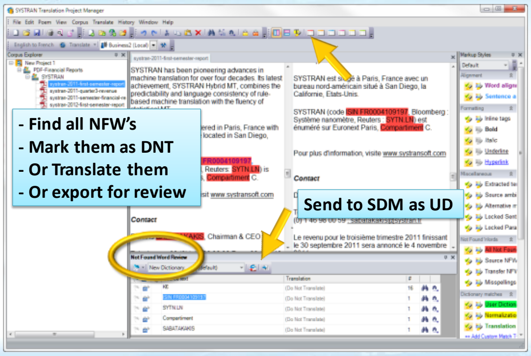
Website & File Directory Crawler


SYSTRAN Translation Project Manager
File Edit Poem View Corpus Translate History Window Help
Corpus Explorer
New Project 2
Processing Status
Crawling

6 Types of Reviews

- NFW Review
  - DNT or Multi lingual entry
  - Export to csv
  - Send to SDM
- Extracted Terms
- Source Ambiguity
- Alternative Meaning
- Sentence
- Translation Memory

Not Found Words (NFW) Review



systran-2011-first-semester-report

SYSTRAN has been pioneering advances in machine translation for over four decades. Its latest achievement, SYSTRAN Hybrid MT, combines the predictability and language consistency of rule-based machine translation with the fluency of

SYSTRAN est siégé à Paris, France avec un bureau nord-américain situé à San Diego, la Californie, Etats-Unis.

SYSTRAN (code ISIN FR0004109197, Bloomberg : Système nanomètre, Reuters : SYTN.LN) est énuméré sur Euronext Paris, Compartiment C.

Pour plus d'information, visite www.systransoft.com

Contact

Chairman & CEO

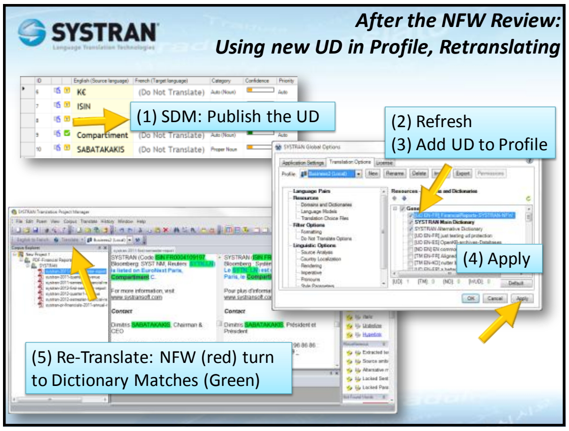
(0) 1 46 98 00 59 : sabatakakis@systran.fr

Le revenu pour le troisième trimestre 2011 finissant le 30 septembre 2011 sera annoncé le 4 novembre

Source text	Translation	#
KE	(Do Not Translate)	16
ISIN FR0004109197	(Do Not Translate)	1
SYTN.LN	(Do Not Translate)	1
Compartiment	(Do Not Translate)	1
SABATAKAKIS	(Do Not Translate)	1

- Find all NFW's
- Mark them as DNT
- Or Translate them
- Or export for review

Send to SDM as UD



After the NFW Review: Using new UD in Profile, Retranslating

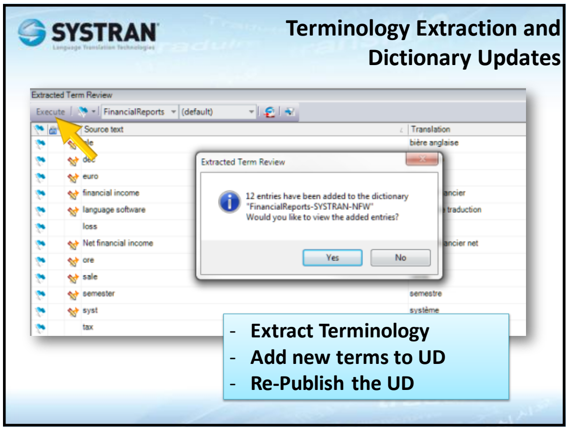
(1) SDM: Publish the UD

(2) Refresh

(3) Add UD to Profile

(4) Apply

(5) Re-Translate: NFW (red) turn to Dictionary Matches (Green)



Terminology Extraction and Dictionary Updates

Extracted Term Review

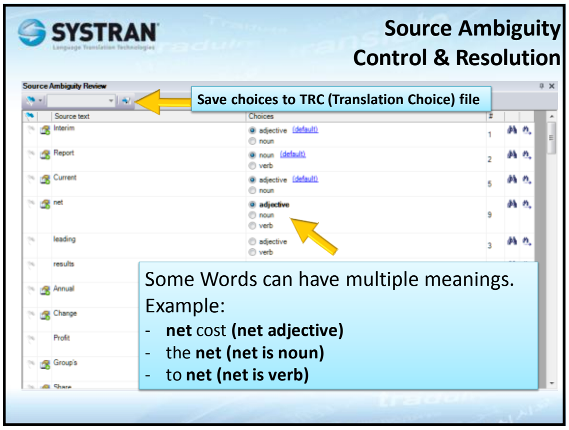
Extracted Term Review

12 entries have been added to the dictionary "FinancialReports-SYSTRAN-NFW"
Would you like to view the added entries?

Yes No

Source text	Translation
ale	bière anglaise
dec	
euro	
financial income	ancier
language software	traduction
loss	
Net financial income	ancier net
ore	
sale	
semester	semestre
syst	système
tax	

- Extract Terminology
- Add new terms to UD
- Re-Publish the UD



Source Ambiguity Control & Resolution

Source Ambiguity Review

Save choices to TRC (Translation Choice) file

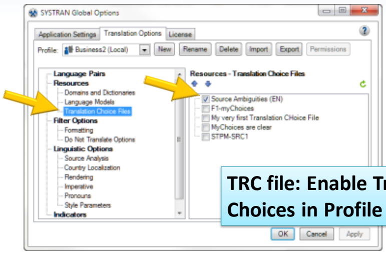
Source text	Choices	#
Interim	● adjective (default) ○ noun	1
Report	● noun (default) ○ verb	2
Current	● adjective (default) ○ noun	5
net	● **adjective** ○ noun ○ verb	9
leading	○ adjective ○ verb	3
results		
Annual		
Change		
Profit		
Group's		
Share		

Some Words can have multiple meanings.
Example:

- **net** cost **(net adjective)**
- the **net (net is noun)**
- to **net (net is verb)**

SYSTRAN Language Translation Technologies

TRC File activation in Profile



SYSTRAN Global Options

Application Settings | Translation Options | License

Profile: Business2 (Local)

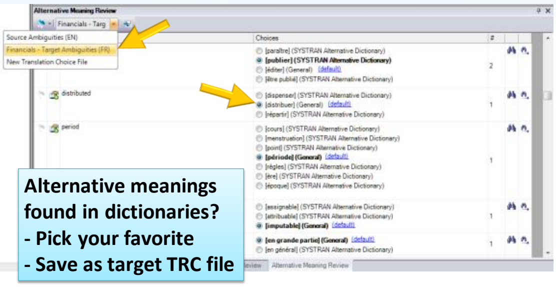
Resources - Translation Choice Files

☑ Source Ambiguities (EN)
☐ F1-myChoices
☐ My very first Translation CHoice File
☐ MyChoices are clear
☐ STPM-SRC1

TRC file: Enable Translation Choices in Profile

SYSTRAN Language Translation Technologies

Target Ambiguity (Alternative Meanings) Control & Resolution



Alternative Meaning Review

Financials - Targ

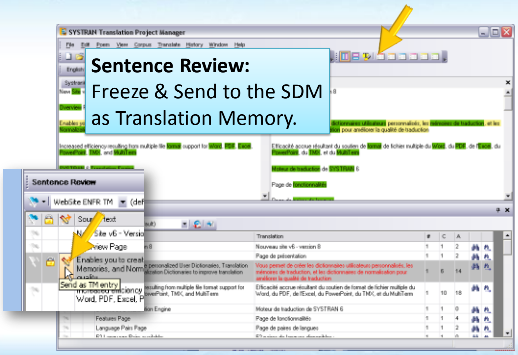
Source Ambiguities (EN)
Financials - Target Ambiguities (FR)
New Translation Choice File

	Choices	#
	○ [paraître] (SYSTRAN Alternative Dictionary) ● [publier] (SYSTRAN Alternative Dictionary) ○ [éditer] (General) (default) ○ [être publié] (SYSTRAN Alternative Dictionary)	2
distributed	○ [dispenser] (SYSTRAN Alternative Dictionary) ● [distribuer] (General) (default) ○ [répartir] (SYSTRAN Alternative Dictionary)	1
period	○ [cours] (SYSTRAN Alternative Dictionary) ○ [menstruation] (SYSTRAN Alternative Dictionary) ○ [point] (SYSTRAN Alternative Dictionary) ● [période] (General) (default) ○ [règles] (SYSTRAN Alternative Dictionary) ○ [ère] (SYSTRAN Alternative Dictionary) ○ [époque] (SYSTRAN Alternative Dictionary)	1
	○ [assignable] (SYSTRAN Alternative Dictionary) ○ [attribuable] (SYSTRAN Alternative Dictionary) ● [imputable] (General) (default)	1
	● [en grande partie] (General) (default) ○ [en général] (SYSTRAN Alternative Dictionary)	1

Alternative meanings found in dictionaries?
- Pick your favorite
- Save as target TRC file

SYSTRAN Language Translation Technologies

Translation Memory Creation

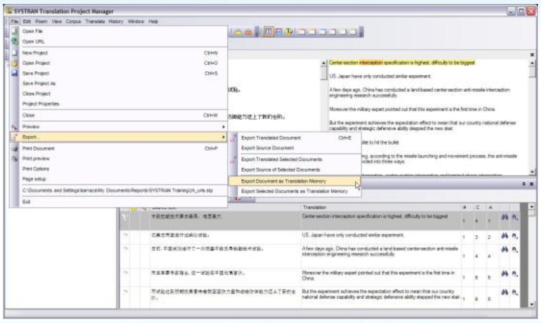


SYSTRAN Translation Project Manager
File Edit Poem View Corpus Translate History Window Help
Sentence Review
WebSite ENFR TM
Send as TM entry
Translation
Features Page
Language Pairs Page

Sentence Review:
Freeze & Send to the SDM as Translation Memory.

SYSTRAN Language Translation Technologies

Exporting a Translation as TM


SYSTRAN Translation Project Manager
File Edit Poem View Corpus Translate History Window Help
Open File
Open URL
New Project Ctrl+N
Open Project Ctrl+O
Save Project Ctrl+S
Save Project As
Close Project
Project Properties
Close Ctrl+W
Preview
Export...
Print Document Ctrl+P
Print preview
Print Options
Page setup
C:\Documents and Settings\barrazaI\My Documents\Reports\SYSTRAN Training\zh_urls.stp
Exit
Export Translated Document Ctrl+E
Export Source Document
Export Translated Selected Documents
Export Source of Selected Documents
Export Document as Translation Memory
Export Selected Documents as Translation Memory
Translation
#
C
A

Center-section interception specification is highest, difficulty to be biggest

US. Japan have only conducted similar experiment.

A few days ago. China has conducted a land-based center-section anti-missile interception engineering research successfully.

Moreover the military expert pointed out that this experiment is the first time in China.

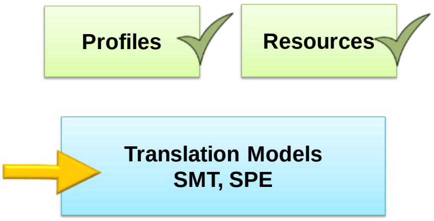
But the experiment achieves the expectation effect to mean that our country national defense capability and strategic defensive ability stepped the new stair

SYSTRAN Language Translation Technologies

What's Next?

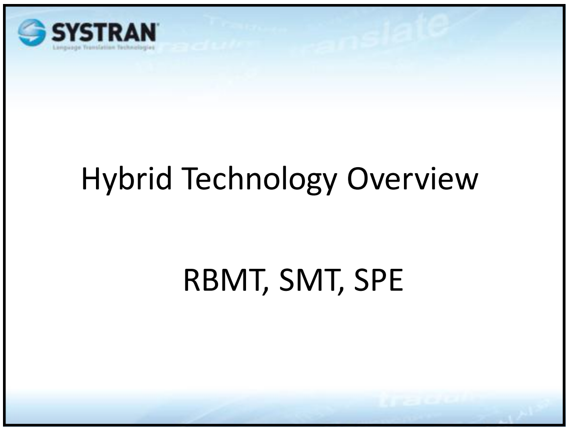
SMT & Hybrid Translation Models

SYSTRAN Language Translation Technologies


Profiles ✓

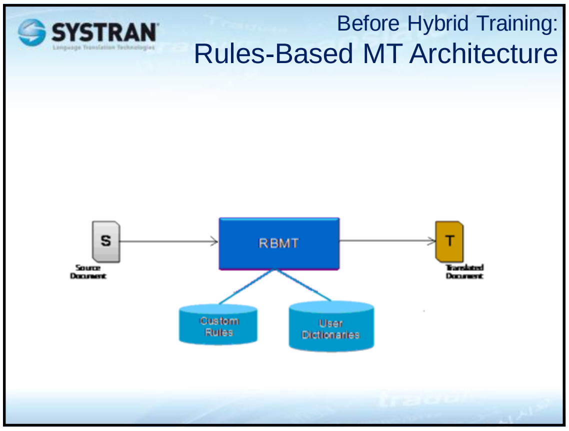
Resources ✓

Translation Models
SMT, SPE



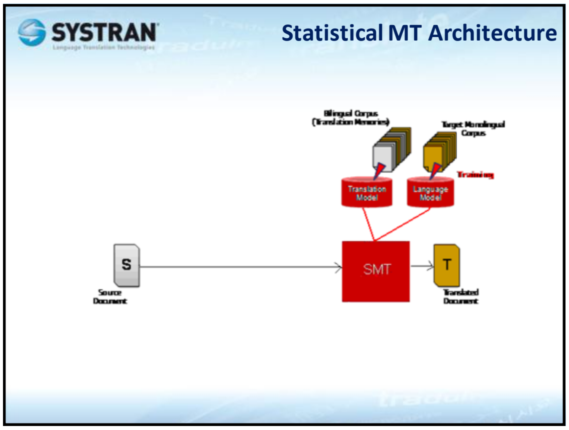
Hybrid Technology Overview

RBMT, SMT, SPE

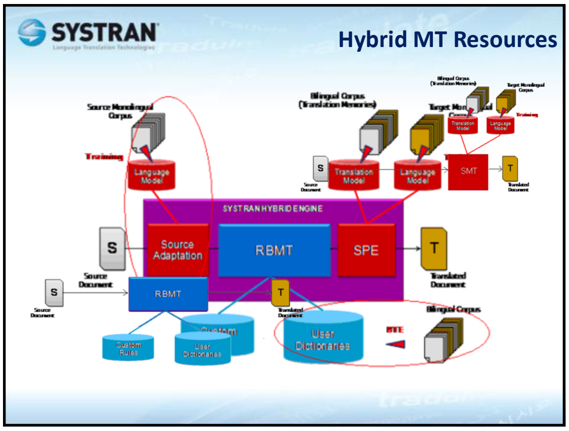


Before Hybrid Training:

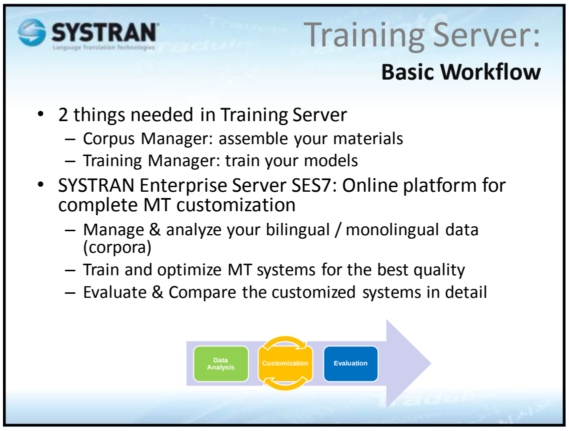
Rules-Based MT Architecture



Statistical MT Architecture



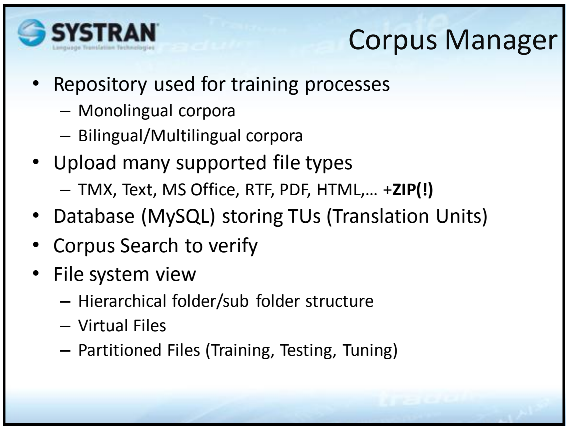
Hybrid MT Resources



Training Server:
Basic Workflow

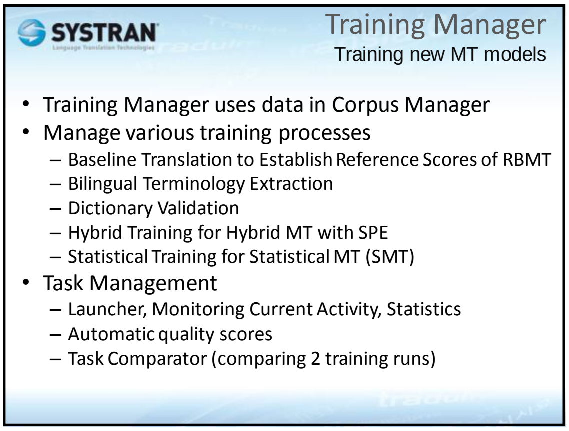
- 2 things needed in Training Server
  - Corpus Manager: assemble your materials
  - Training Manager: train your models
- SYSTRAN Enterprise Server SES7: Online platform for complete MT customization
  - Manage & analyze your bilingual / monolingual data (corpora)
  - Train and optimize MT systems for the best quality
  - Evaluate & Compare the customized systems in detail

SYSTRAN
Language Translation Technologies
Data Analysis
Customization
Evaluation



Corpus Manager

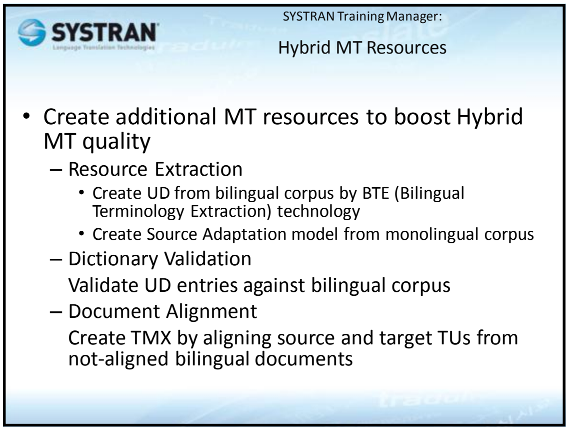
- Repository used for training processes
  - Monolingual corpora
  - Bilingual/Multilingual corpora
- Upload many supported file types
  - TMX, Text, MS Office, RTF, PDF, HTML,… +**ZIP(!)**
- Database (MySQL) storing TUs (Translation Units)
- Corpus Search to verify
- File system view
  - Hierarchical folder/sub folder structure
  - Virtual Files
  - Partitioned Files (Training, Testing, Tuning)



Training Manager

Training new MT models

- Training Manager uses data in Corpus Manager
- Manage various training processes
  - Baseline Translation to Establish Reference Scores of RBMT
  - Bilingual Terminology Extraction
  - Dictionary Validation
  - Hybrid Training for Hybrid MT with SPE
  - Statistical Training for Statistical MT (SMT)
- Task Management
  - Launcher, Monitoring Current Activity, Statistics
  - Automatic quality scores
  - Task Comparator (comparing 2 training runs)



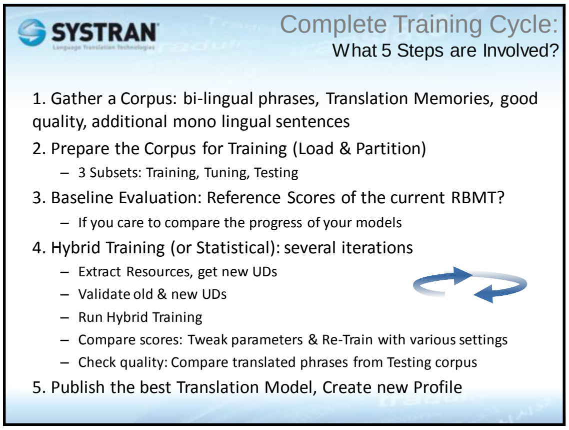
SYSTRAN Training Manager:

Hybrid MT Resources

- Create additional MT resources to boost Hybrid MT quality
  - Resource Extraction
    - Create UD from bilingual corpus by BTE (Bilingual Terminology Extraction) technology
    - Create Source Adaptation model from monolingual corpus
  - Dictionary Validation

    Validate UD entries against bilingual corpus
  - Document Alignment

    Create TMX by aligning source and target TUs from not-aligned bilingual documents

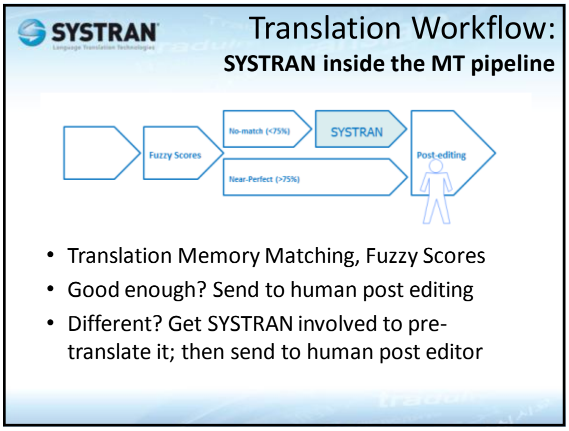


Complete Training Cycle:

What 5 Steps are Involved?

1. Gather a Corpus: bi-lingual phrases, Translation Memories, good quality, additional mono lingual sentences
2. Prepare the Corpus for Training (Load & Partition)
   - 3 Subsets: Training, Tuning, Testing
3. Baseline Evaluation: Reference Scores of the current RBMT?
   - If you care to compare the progress of your models
4. Hybrid Training (or Statistical): several iterations
   - Extract Resources, get new UDs
   - Validate old & new UDs
   - Run Hybrid Training
   - Compare scores: Tweak parameters & Re-Train with various settings
   - Check quality: Compare translated phrases from Testing corpus
5. Publish the best Translation Model, Create new Profile

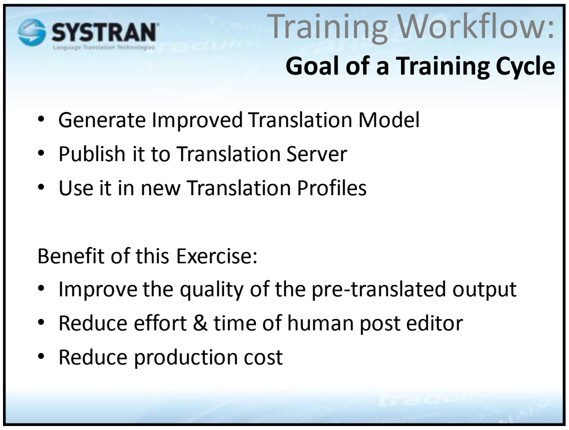
Translation Workflow
&
Training Workflow



Translation Workflow:

SYSTRAN inside the MT pipeline

- Translation Memory Matching, Fuzzy Scores
- Good enough? Send to human post editing
- Different? Get SYSTRAN involved to pre-translate it; then send to human post editor



Training Workflow:

Goal of a Training Cycle

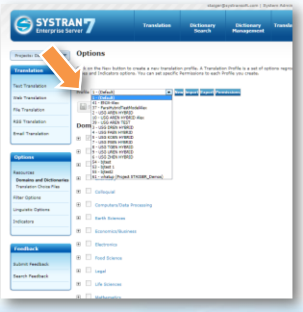
- Generate Improved Translation Model
- Publish it to Translation Server
- Use it in new Translation Profiles

Benefit of this Exercise:

- Improve the quality of the pre-translated output
- Reduce effort & time of human post editor
- Reduce production cost

The Ultimate Goal:

Smarter Profiles, Better Translation Models



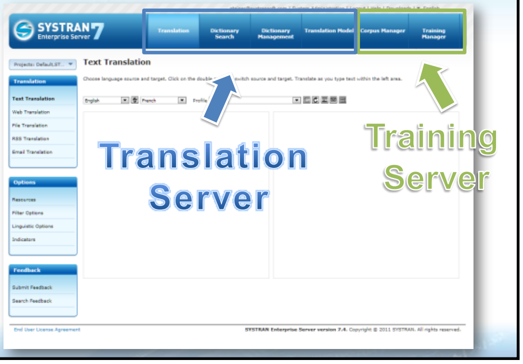
- Enhance the Rules Based System, e.g. with new UDs
- Enhance the Statistical Post Editing (SPE) with bilingual Translation Models
- Unsupported language? Create purely statistical model when there's no rules engine
- Smarter Technology, Better Translation

Corpus Manager

Step-by-step

After you login

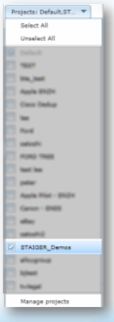
The Full Interface



SYSTRAN Language Translation Technologies

Another Reminder:

Working within a Project



→Select a Project

→Or create a new one

SYSTRAN Language Translation Technologies

Reminders:

What we're about to do

→ **Gather the Corpus**
- Bilingual files (TMX, BiText…)
  - Optionally additional monolingual files
- Partitioning: Getting it ready for Training
  - 3 parts: Training corpus, Tuning corpus, testing corpus

→ **Train a new Translation System**
- Measure (Score) the current system: Baseline as Reference
- Train, several iterations as needed:
  - Run Training (Hybrid, or purely Statistical)
  - Compare the Scores
  - Extract Resources, Validate User Dictionaries
  - Tweak parameters & re-Tune / re-Train as Needed

→Publish the new resources

SYSTRAN Language Translation Technologies

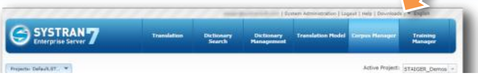
Corpus Manager step-by-step

Reminder: BiText Format

Easily create your own TM

```
#ENCODING=UTF-8
#TM                                   🡐 this is a Translation Memory!
#EN<TAB>ES                            🡐 use a 'TAB' to separate EN and  ES
First sentence.<TAB>Primera frase.
Got a second phrase.<TAB>Tengo una Segunda frase.
Three is a charm.<TAB>Tres son muy buenos.
```

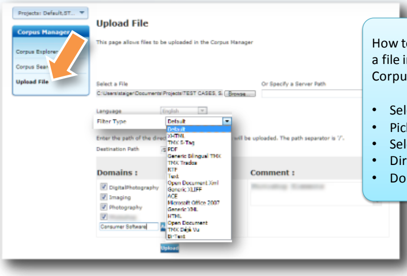
Corpus Explorer



Active Project: STAIGER_Demos

Corpus Manager

Uploading a File



Upload File

This page allows files to be uploaded in the Corpus Manager

Select a File: C:\Users\staiger\Documents\Projects\TEST CASES, S

Or Specify a Server Path

Language: English

Filter Type: Default

Enter the path of the direc... will be uploaded. The path separator is '/'.

Destination Path

Domains :

☑ DigitalPhotography

☑ Imaging

☑ Photography

☑

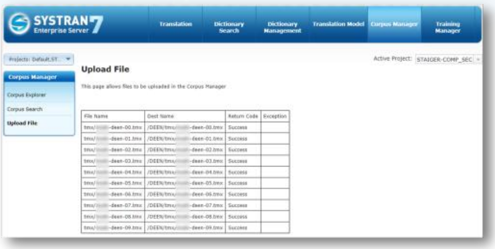
Consumer Software

How to upload a file into the Corpus Manager:

- Select File
- Pick Source Language
- Select Filter type
- Directory
- Domains & Comment

Upload Example

Multiple files from a ZIP archive



Active Project: STAIGER-COMP_SEC

Upload File

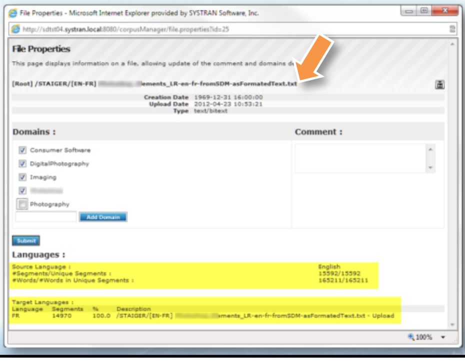
This page allows files to be uploaded in the Corpus Manager

File Name	Dest Name	Return Code	Exception
tmx/[redacted]-deen-00.tmx	/DEEN/tmx/[redacted]-deen-00.tmx	Success	
tmx/[redacted]-deen-01.tmx	/DEEN/tmx/[redacted]-deen-01.tmx	Success	
tmx/[redacted]-deen-02.tmx	/DEEN/tmx/[redacted]-deen-02.tmx	Success	
tmx/[redacted]-deen-03.tmx	/DEEN/tmx/[redacted]-deen-03.tmx	Success	
tmx/[redacted]-deen-04.tmx	/DEEN/tmx/[redacted]-deen-04.tmx	Success	
tmx/[redacted]-deen-05.tmx	/DEEN/tmx/[redacted]-deen-05.tmx	Success	
tmx/[redacted]-deen-06.tmx	/DEEN/tmx/[redacted]-deen-06.tmx	Success	
tmx/[redacted]-deen-07.tmx	/DEEN/tmx/[redacted]-deen-07.tmx	Success	
tmx/[redacted]-deen-08.tmx	/DEEN/tmx/[redacted]-deen-08.tmx	Success	
tmx/[redacted]-deen-09.tmx	/DEEN/tmx/[redacted]-deen-09.tmx	Success	

➔ Got Many files? Upload a Zip archive of the files

Corpus Explorer:

File Properties


File Properties - Microsoft Internet Explorer provided by SYSTRAN Software, Inc.
100%

http://sdtst04.systran.local:8080/corpusManager/file.properties?id=25

File Properties

This page displays information on a file, allowing update of the comment and domains d

[Root] /STAIGER/[EN-FR] [redacted]ements_LR-en-fr-fromSDM-asFormatedText.txt

Creation Date 1969-12-31 16:00:00
Upload Date 2012-04-23 10:53:21
Type text/bitext

Domains :

☑ Consumer Software
☑ DigitalPhotography
☑ Imaging
☑ [redacted]
☐ Photography

Add Domain

Comment :

Submit

Languages :

Source Language :	English
#Segments/Unique Segments :	15592/15592
#Words/#Words in Unique Segments :	165211/165211

Target Languages :

Language	Segments	%	Description
FR	14970	100.0	/STAIGER/[EN-FR] [redacted]ements_LR-en-fr-fromSDM-asFormatedText.txt - Upload

Combining multiple files

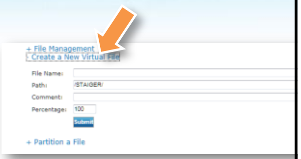
Virtual Files

Got (too) many files?

➔Wrap them into a single "virtual" file

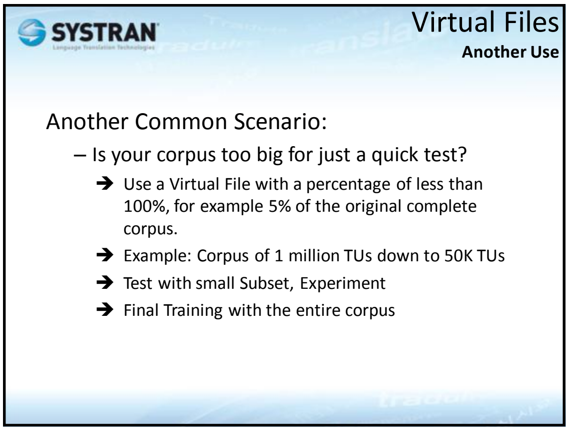
➔Even if from different file types

➔Treat it as a single entity for Training purposes



File Name:
Path: /STAIGER/
Comment:
Percentage: 100

1. Select the files in the explorer view
2. Click 'Create a New Virtual File'
3. File Name: give your virtual file a name
4. Path: set Directory location for the virtual file
5. Add an optional comment
6. Indicate if you want all of their content (100%) or just a (randomly selected) subset
7. Go!

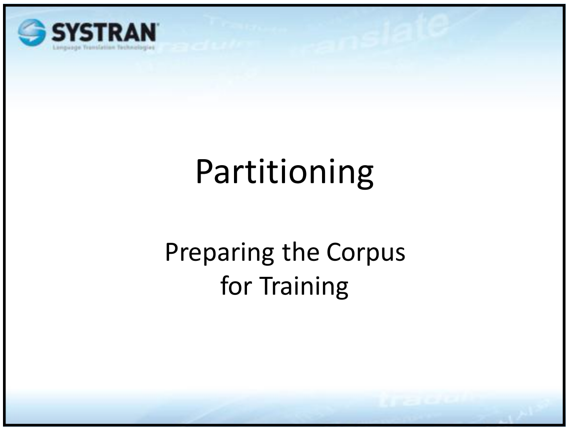


Virtual Files

Another Use

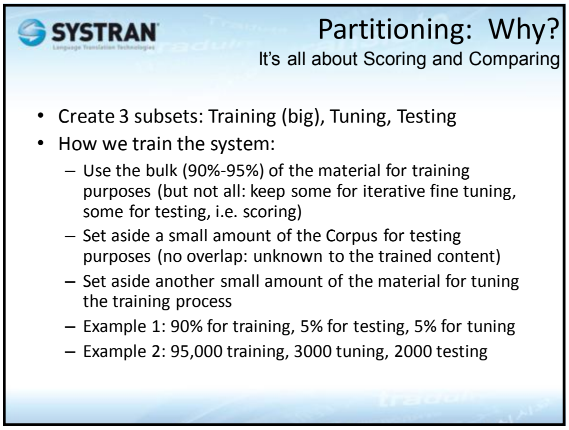
Another Common Scenario:

- Is your corpus too big for just a quick test?
  - ➔ Use a Virtual File with a percentage of less than 100%, for example 5% of the original complete corpus.
  - ➔ Example: Corpus of 1 million TUs down to 50K TUs
  - ➔ Test with small Subset, Experiment
  - ➔ Final Training with the entire corpus



Partitioning

Preparing the Corpus
for Training



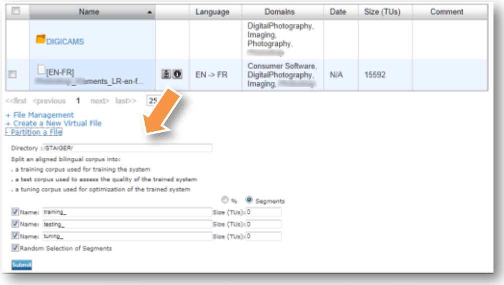
Partitioning: Why?

It's all about Scoring and Comparing

- Create 3 subsets: Training (big), Tuning, Testing
- How we train the system:
  - Use the bulk (90%-95%) of the material for training purposes (but not all: keep some for iterative fine tuning, some for testing, i.e. scoring)
  - Set aside a small amount of the Corpus for testing purposes (no overlap: unknown to the trained content)
  - Set aside another small amount of the material for tuning the training process
  - Example 1: 90% for training, 5% for testing, 5% for tuning
  - Example 2: 95,000 training, 3000 tuning, 2000 testing

Partitioning: How?

Using the Web Interface



☐	Name ▲		Language	Domains	Date	Size (TUs)	Comment
	DIGICAMS			DigitalPhotography, Imaging, Photography, [redacted]			
☐	[EN-FR] [redacted]_[redacted]ements_LR-en-f...		EN -> FR	Consumer Software, DigitalPhotography, Imaging, [redacted]	N/A	15592	

<<first <previous 1 next> last>> 25

+ File Management
+ Create a New Virtual File
- Partition a File

Directory : /STAIGER/

Split an aligned bilingual corpus into:
. a training corpus used for training the system
. a test corpus used to assess the quality of the trained system
. a tuning corpus used for optimization of the trained system

○ % ● Segments

☑ Name: training_ Size (TUs): 0
☑ Name: testing_ Size (TUs): 0
☑ Name: tuning_ Size (TUs): 0
☑ Random Selection of Segments

Partitioning

Are there Alternatives?

Use handy Web interface to partition, or:

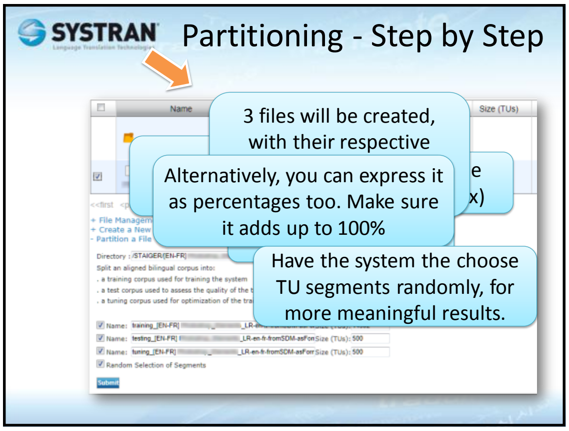
- Create 3 files and upload them into a folder:
  - training_corpus.txt
  - testing_corpus.txt
  - tuning_corpus.txt
- Have additional monolingual corpora?
  - mono_EN.txt
  - mono_FR.txt
  - mono_DE.txt
  - etc…

Partitioning

If you create your own

Avoid Overlaps – don't cheat yourself

- Segments in one corpus should not exist in any of the other two corpora:
  - Training, tuning, testing
  - You don't want to run tests on what's already been seen in the training



Partitioning - Step by Step

3 files will be created, with their respective

e x)

Alternatively, you can express it as percentages too. Make sure it adds up to 100%

Have the system the choose TU segments randomly, for more meaningful results.

Directory : /STAIGER/[EN-FR] [redacted]

Split an aligned bilingual corpus into:

. a training corpus used for training the system

. a test corpus used to assess the quality of the t

. a tuning corpus used for optimization of the tra

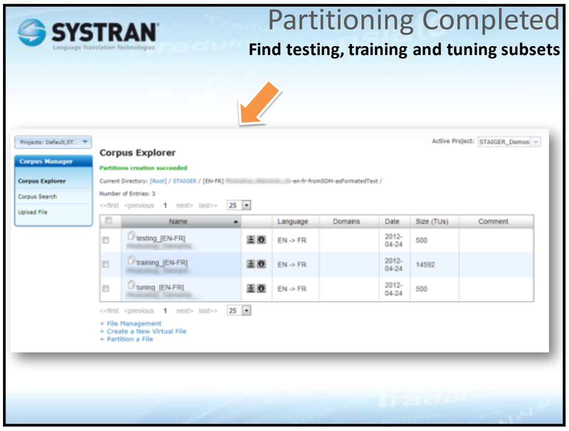
☑ Name: training_[EN-FR] [redacted] _LR-en-

☑ Name: testing_[EN-FR] [redacted] _LR-en-fr-fromSDM-asFon Size (TUs): 500

☑ Name: tuning_[EN-FR] [redacted] _LR-en-fr-fromSDM-asForr Size (TUs): 500

☑ Random Selection of Segments

Submit



Partitioning Completed

Find testing, training and tuning subsets

SYSTRAN
Language Translation Technologies
Projects: Default,ST...
Active Project: STAIGER_Demos
Corpus Manager
Corpus Explorer
Corpus Search
Upload File
<<first <previous 1 next> last>> 25
+ File Management
+ Create a New Virtual File
+ Partition a File

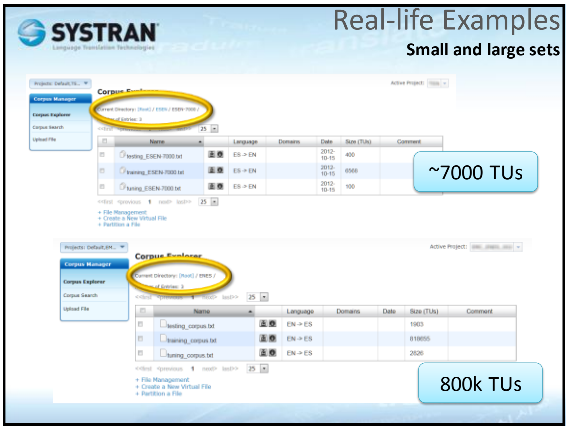
Corpus Explorer

Partitions creation succeeded

Current Directory: [Root] / STAIGER / [EN-FR] [redacted]-en-fr-fromSDM-asFormatedText /

Number of Entries: 3

□	Name		Language	Domains	Date	Size (TUs)	Comment
□	testing_[EN-FR] [redacted]		EN -> FR		2012-04-24	500	
□	training_[EN-FR] [redacted]		EN -> FR		2012-04-24	14592	
□	tuning_[EN-FR] [redacted]		EN -> FR		2012-04-24	500	



Real-life Examples

Small and large sets

Corpus Explorer

Current Directory: [Root] / ESEN / ESEN-7000 /

Number of Entries: 3

□	Name		Language	Domains	Date	Size (TUs)	Comment
□	testing_ESEN-7000.txt		ES -> EN		2012-10-15	400	
□	training_ESEN-7000.txt		ES -> EN		2012-10-15	6568	
□	tuning_ESEN-7000.txt		ES -> EN		2012-10-15	100	

~7000 TUs

Corpus Explorer

Current Directory: [Root] / ENES /

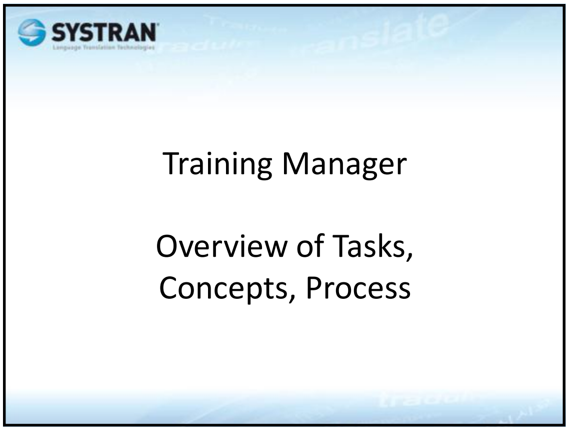
Number of Entries: 3

□	Name		Language	Domains	Date	Size (TUs)	Comment
□	testing_corpus.txt		EN -> ES			1903	
□	training_corpus.txt		EN -> ES			818655	
□	tuning_corpus.txt		EN -> ES			2826	

800k TUs

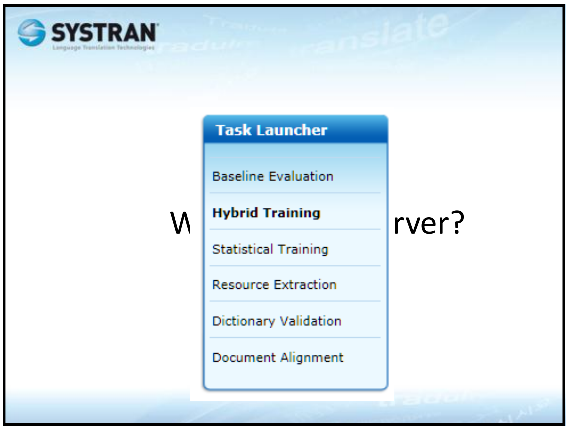
Ready to use the corpus
for training purposes

Next: Training Manager

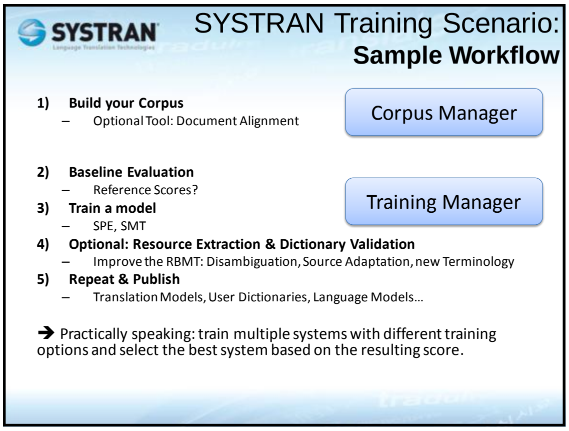


Training Manager

Overview of Tasks,
Concepts, Process



W rver?



SYSTRAN Training Scenario: **Sample Workflow**

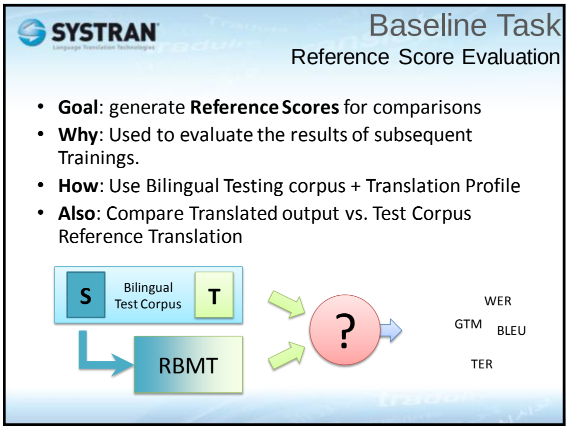
1) **Build your Corpus**
   - Optional Tool: Document Alignment
2) **Baseline Evaluation**
   - Reference Scores?
3) **Train a model**
   - SPE, SMT
4) **Optional: Resource Extraction & Dictionary Validation**
   - Improve the RBMT: Disambiguation, Source Adaptation, new Terminology
5) **Repeat & Publish**
   - Translation Models, User Dictionaries, Language Models…

Corpus Manager
Training Manager

➔ Practically speaking: train multiple systems with different training options and select the best system based on the resulting score.

Training Manager:

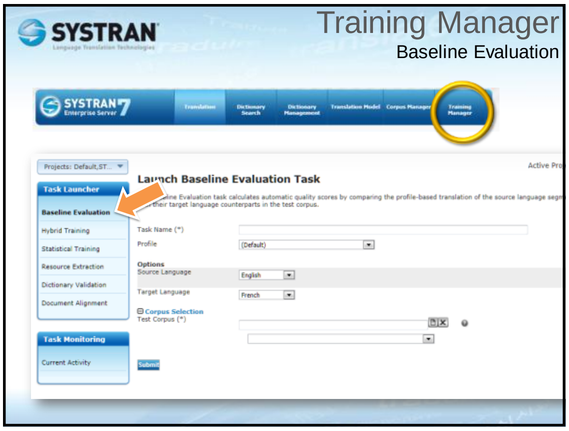
Baseline Evaluation



Baseline Task
Reference Score Evaluation

- **Goal**: generate **Reference Scores** for comparisons
- **Why**: Used to evaluate the results of subsequent Trainings.
- **How**: Use Bilingual Testing corpus + Translation Profile
- **Also**: Compare Translated output vs. Test Corpus Reference Translation

S
Bilingual Test Corpus
T
RBMT
?
WER
GTM
BLEU
TER



Training Manager

Baseline Evaluation

Launch Baseline Evaluation Task

...line Evaluation task calculates automatic quality scores by comparing the profile-based translation of the source language segm... ...their target language counterparts in the test corpus.

Task Name (*)

Profile: (Default)

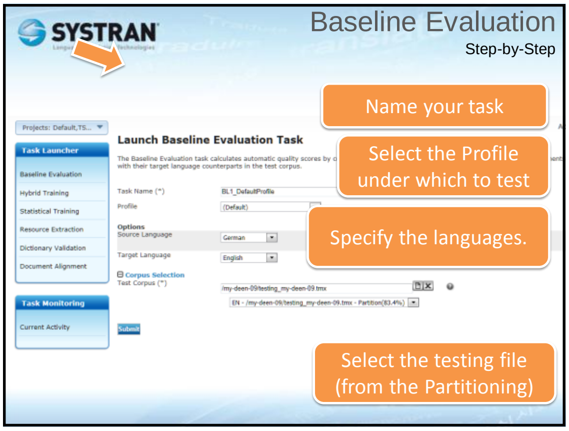
Options

Source Language: English

Target Language: French

Corpus Selection

Test Corpus (*)



Baseline Evaluation

Step-by-Step

Name your task

Select the Profile under which to test

Specify the languages.

Select the testing file (from the Partitioning)

Launch Baseline Evaluation Task

The Baseline Evaluation task calculates automatic quality scores by c... with their target language counterparts in the test corpus.

Task Name (*): BL1_DefaultProfile

Profile: (Default)

Options

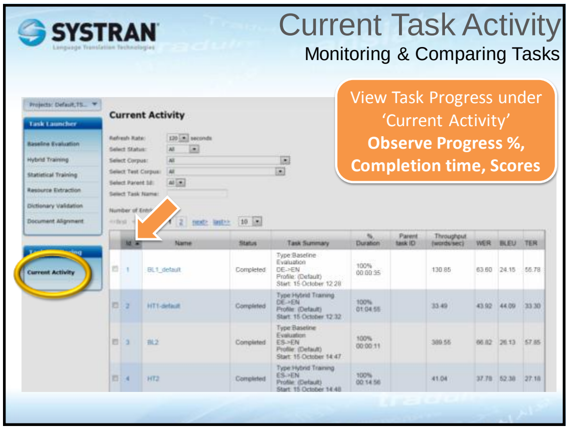
Source Language: German

Target Language: English

Corpus Selection

Test Corpus (*): /my-deen-09/testing_my-deen-09.tmx

EN - /my-deen-09/testing_my-deen-09.tmx - Partition(83.4%)



Current Task Activity

Monitoring & Comparing Tasks

View Task Progress under ‘Current Activity’
Observe Progress %, Completion time, Scores

Current Activity

Refresh Rate: 120 seconds

Select Status: All

Select Corpus: All

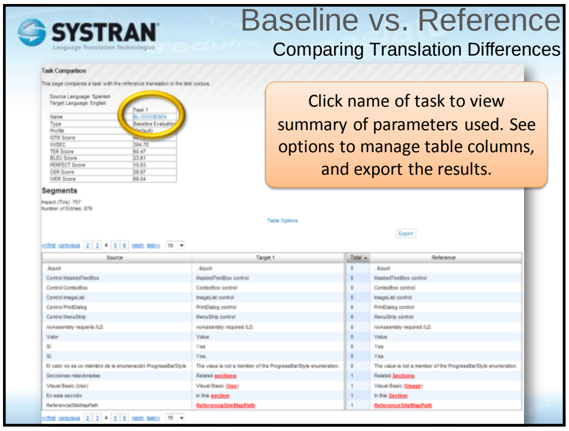
Select Test Corpus: All

Select Parent Id: All

Select Task Name:

Number of Entri...

	Id	Name	Status	Task Summary	% Duration	Parent task ID	Throughput (words/sec)	WER	BLEU	TER
□	1	BL1_default	Completed	Type:Baseline Evaluation DE->EN Profile: (Default) Start: 15 October 12:28	100% 00:00:35		130.85	63.60	24.15	55.78
□	2	HT1-default	Completed	Type:Hybrid Training DE->EN Profile: (Default) Start: 15 October 12:32	100% 01:04:55		33.49	43.92	44.09	33.30
□	3	BL2	Completed	Type:Baseline Evaluation ES->EN Profile: (Default) Start: 15 October 14:47	100% 00:00:11		389.55	66.82	26.13	57.85
□	4	HT2	Completed	Type:Hybrid Training ES->EN Profile: (Default) Start: 15 October 14:48	100% 00:14:56		41.04	37.78	52.38	27.18



Baseline vs. Reference

Comparing Translation Differences

Task Comparison

This page compares a task with the reference translation in the test corpus.

Source Language: Spanish
Target Language: English

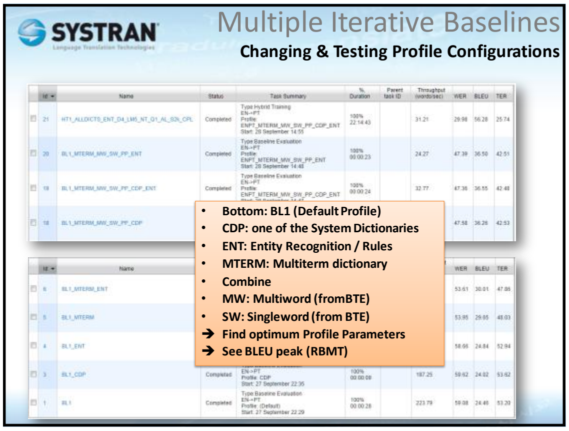
	Task 1
Name	BL-[redacted]ESEN
Type	Baseline Evaluation
Profile	(Default)
GTM Score	[illegible]
WISEC	384.70
TER Score	60.47
BLEU Score	23.61
PERFECT Score	10.63
CER Score	38.87
WER Score	68.04

Segments

Impact (TUs): 757
Number of Entries: 879

Click name of task to view summary of parameters used. See options to manage table columns, and export the results.

Source	Target 1	Total	Reference
,&quot	, &quot	0	, &quot
Control MaskedTextBox	MaskedTextBox control	0	MaskedTextBox control
Control ComboBox	ComboBox control	0	ComboBox control
Control ImageList	ImageList control	0	ImageList control
Control PrintDialog	PrintDialog control	0	PrintDialog control
Control MenuStrip	MenuStrip control	0	MenuStrip control
noAssembly requería /LD.	noAssembly required /LD.	0	noAssembly required /LD.
Valor	Value	0	Value
Sí	Yes	0	Yes
Sí.	Yes.	0	Yes.
El valor no es un miembro de la enumeración ProgressBarStyle.	The value is not a member of the ProgressBarStyle enumeration.	0	The value is not a member of the ProgressBarStyle enumeration.
Secciones relacionadas	Related **sections**	1	Related **Sections**
Visual Basic (Uso)	Visual Basic (**Use**)	1	Visual Basic (**Usage**)
En esta sección	In this **section**	1	In this **Section**
ReferenciaSiteMapPath	**ReferenciaSiteMapPath**	1	**ReferenceSiteMapPath**



Multiple Iterative Baselines

Changing & Testing Profile Configurations

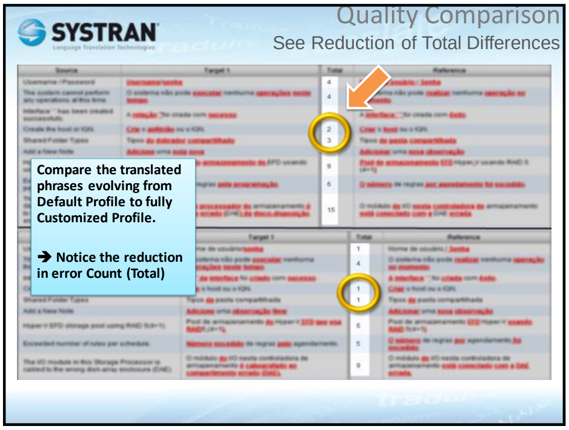
Id	Name	Status	Task Summary	% Duration	Parent task ID	Throughput (words/sec)	WER	BLEU	TER
21	HT1_ALLDICTS_ENT_D4_LM5_NT_Q1_AL_S2k_CPL	Completed	Type Hybrid Training EN->PT Profile: ENPT_MTERM_MW_SW_PP_CDP_ENT Start: 28 September 14:55	100% 22:14:43		31.21	29.98	56.28	25.74
20	BL1_MTERM_MW_SW_PP_ENT	Completed	Type Baseline Evaluation EN->PT Profile: ENPT_MTERM_MW_SW_PP_ENT Start: 28 September 14:48	100% 00:00:23		24.27	47.39	36.50	42.51
19	BL1_MTERM_MW_SW_PP_CDP_ENT	Completed	Type Baseline Evaluation EN->PT Profile: ENPT_MTERM_MW_SW_PP_CDP_ENT	100% 00:00:24		32.77	47.36	36.55	42.48
18	BL1_MTERM_MW_SW_PP_CDP						47.58	36.26	42.53

Id	Name	Status	Task Summary	% Duration	Parent task ID	Throughput (words/sec)	WER	BLEU	TER
6	BL1_MTERM_ENT						53.61	30.01	47.86
5	BL1_MTERM						53.95	29.65	48.03
4	BL1_ENT						58.66	24.84	52.94
3	BL1_CDP	Completed	EN->PT Profile: CDP Start: 27 September 22:35	100% 00:00:08		187.25	59.62	24.02	53.62
1	BL1	Completed	Type Baseline Evaluation EN->PT Profile: (Default) Start: 27 September 22:29	100% 00:00:28		223.79	59.08	24.46	53.20

- **Bottom: BL1 (Default Profile)**
- **CDP: one of the System Dictionaries**
- **ENT: Entity Recognition / Rules**
- **MTERM: Multiterm dictionary**
- **Combine**
- **MW: Multiword (fromBTE)**
- **SW: Singleword (from BTE)**

➔ **Find optimum Profile Parameters**

➔ **See BLEU peak (RBMT)**



Quality Comparison

See Reduction of Total Differences

Compare the translated phrases evolving from Default Profile to fully Customized Profile.

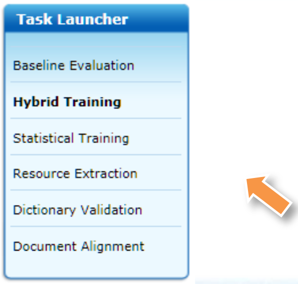
➔ Notice the reduction in error Count (Total)

Next Steps?

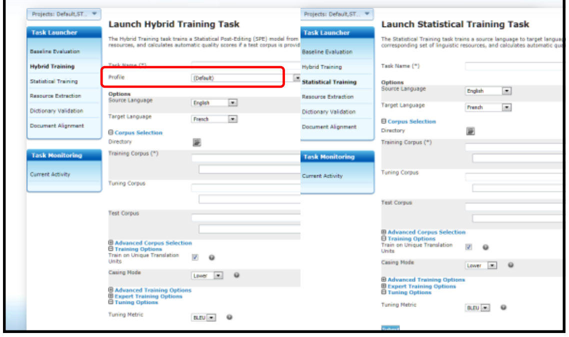
- We've run a Baseline eval (RBMT Reference)
  - Testing corpus
- We have two more Corpora yet unused:
  - Training corpus
  - Tuning corpus
- Ready to train… but which way?
  - Consider & Choose? Maybe both
    - Hybrid (Rules & SPE) translation?
    - Purely Statistical translation?

Two Types of Training

Hybrid vs. Statistical Training



Hybrid vs. Statistical Training



Launch Hybrid Training Task

The Hybrid Training task trains a Statistical Post-Editing (SPE) model from resources, and calculates automatic quality scores if a test corpus is provid

Task Name (*)

Profile (Default)

Options

Source Language English

Target Language French

⊟ Corpus Selection

Directory

Training Corpus (*)

Tuning Corpus

Test Corpus

⊞ Advanced Corpus Selection

⊟ Training Options

Train on Unique Translation Units ☑

Casing Mode Lower

⊞ Advanced Training Options

⊞ Expert Training Options

⊟ Tuning Options

Tuning Metric BLEU

Launch Statistical Training Task

The Statistical Training task trains a source language to target langua corresponding set of linguistic resources, and calculates automatic qu

Task Name (*)

Options

Source Language English

Target Language French

⊟ Corpus Selection

Directory

Training Corpus (*)

Tuning Corpus

Test Corpus

⊞ Advanced Corpus Selection

⊟ Training Options

Train on Unique Translation Units ☑

Casing Mode Lower

⊞ Advanced Training Options

⊞ Expert Training Options

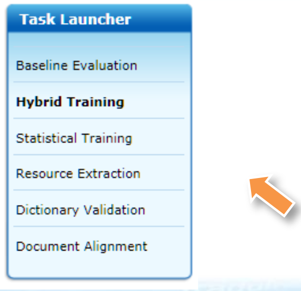
⊟ Tuning Options

Tuning Metric BLEU

SYSTRAN Language Translation Technologies

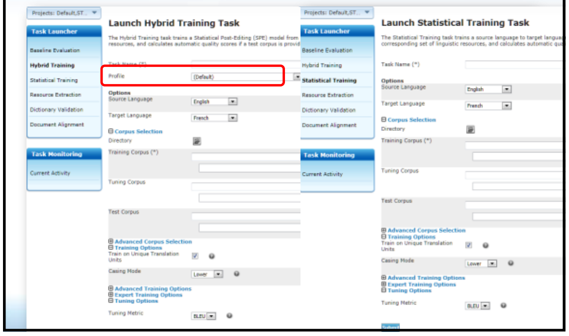
Next:

Two Types of Training



SYSTRAN Language Translation Technologies

Hybrid vs. Statistical Training



Launch Hybrid Training Task

The Hybrid Training task trains a Statistical Post-Editing (SPE) model from resources, and calculates automatic quality scores if a test corpus is provid

Task Name (*)

Profile (Default)

Options

Source Language English

Target Language French

⊟ Corpus Selection

Directory

Training Corpus (*)

Tuning Corpus

Test Corpus

⊞ Advanced Corpus Selection

⊟ Training Options

Train on Unique Translation Units ☑

Casing Mode Lower

⊞ Advanced Training Options

⊞ Expert Training Options

⊟ Tuning Options

Tuning Metric BLEU

Launch Statistical Training Task

The Statistical Training task trains a source language to target langua corresponding set of linguistic resources, and calculates automatic qua

Task Name (*)

Options

Source Language English

Target Language French

⊟ Corpus Selection

Directory

Training Corpus (*)

Tuning Corpus

Test Corpus

⊞ Advanced Corpus Selection

⊟ Training Options

Train on Unique Translation Units ☑

Casing Mode Lower

⊞ Advanced Training Options

⊞ Expert Training Options

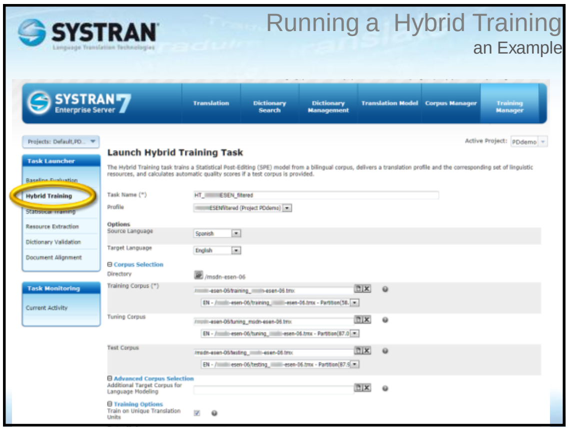
⊟ Tuning Options

Tuning Metric BLEU

SYSTRAN Language Translation Technologies

Training Server

Hybrid Training



Running a Hybrid Training

an Example

Launch Hybrid Training Task

The Hybrid Training task trains a Statistical Post-Editing (SPE) model from a bilingual corpus, delivers a translation profile and the corresponding set of linguistic resources, and calculates automatic quality scores if a test corpus is provided.

Task Name (*): HT_[redacted]ESEN_filtered

Profile: [redacted]ESENfiltered (Project PDdemo)

Options

Source Language: Spanish

Target Language: English

⊟ Corpus Selection

Directory: /msdn-esen-06

Training Corpus (*): [redacted]-esen-06/training_[redacted]-esen-06.tmx
EN - /[redacted]-esen-06/training_[redacted]-esen-06.tmx - Partition(58.

Tuning Corpus: [redacted]-esen-06/tuning_msdn-esen-06.tmx
EN - /[redacted]-esen-06/tuning_[redacted]-esen-06.tmx - Partition(87.0

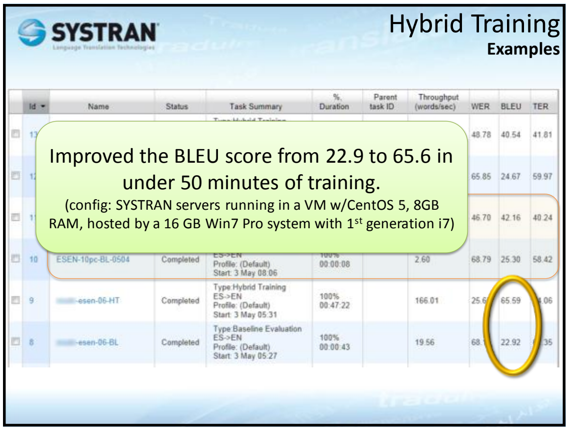
Test Corpus: /msdn-esen-06/testing_[redacted]-esen-06.tmx
EN - /[redacted]-esen-06/testing_[redacted]-esen-06.tmx - Partition(87.9

⊟ Advanced Corpus Selection

Additional Target Corpus for Language Modeling

⊟ Training Options

Train on Unique Translation Units ☑



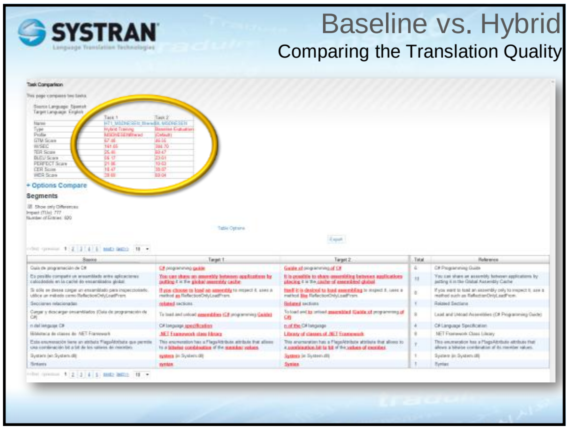
Hybrid Training

Examples

	Id	Name	Status	Task Summary	%, Duration	Parent task ID	Throughput (words/sec)	WER	BLEU	TER
□	13			Type:Hybrid Training				48.78	40.54	41.81
□	12							65.85	24.67	59.97
□	11							46.70	42.16	40.24
□	10	ESEN-10pc-BL-0504	Completed	ES->EN Profile: (Default) Start: 3 May 08:06	100% 00:00:08		2.60	68.79	25.30	58.42
□	9	[redacted]-esen-06-HT	Completed	Type:Hybrid Training ES->EN Profile: (Default) Start: 3 May 05:31	100% 00:47:22		166.01	25.6	65.59	.06
□	8	[redacted]-esen-06-BL	Completed	Type:Baseline Evaluation ES->EN Profile: (Default) Start: 3 May 05:27	100% 00:00:43		19.56	68.	22.92	.35

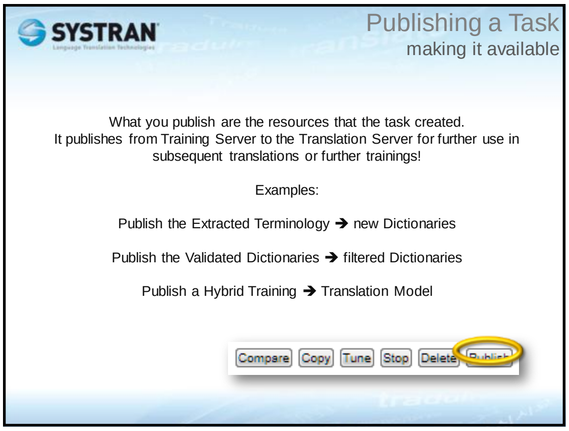
Improved the BLEU score from 22.9 to 65.6 in under 50 minutes of training.

(config: SYSTRAN servers running in a VM w/CentOS 5, 8GB RAM, hosted by a 16 GB Win7 Pro system with 1st generation i7)



Baseline vs. Hybrid

Comparing the Translation Quality



Publishing a Task
making it available

What you publish are the resources that the task created.
It publishes from Training Server to the Translation Server for further use in subsequent translations or further trainings!

Examples:

Publish the Extracted Terminology ➔ new Dictionaries

Publish the Validated Dictionaries ➔ filtered Dictionaries

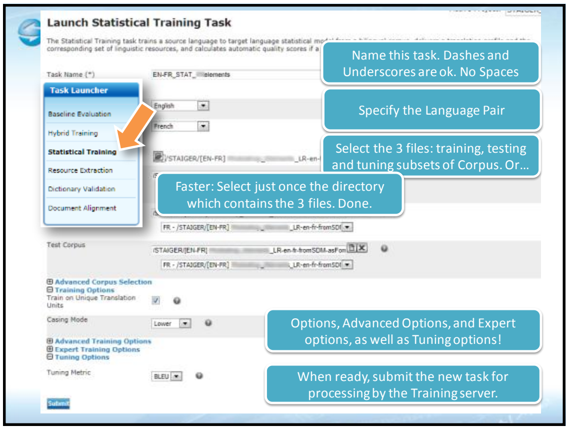
Publish a Hybrid Training ➔ Translation Model

SYSTRAN
Language Translation Technologies
Compare
Copy
Tune
Stop
Delete
Publish

Training Manager:

Statistical Training

SYSTRAN
Language Translation Technologies



Launch Statistical Training Task

The Statistical Training task trains a source language to target language statistical mo… corresponding set of linguistic resources, and calculates automatic quality scores if a …

Task Name (*) EN-FR_STAT_[redacted]elements

English

French

/STAIGER/[EN-FR] [redacted]_LR-en-

FR - /STAIGER/[EN-FR] [redacted]_LR-en-fr-fromSDI

Test Corpus /STAIGER/[EN-FR] [redacted]_LR-en-fr-fromSDM-asFom

FR - /STAIGER/[EN-FR] [redacted]_LR-en-fr-fromSDI

⊞ Advanced Corpus Selection

⊟ Training Options

Train on Unique Translation Units ☑

Casing Mode Lower

⊞ Advanced Training Options

⊞ Expert Training Options

⊟ Tuning Options

Tuning Metric BLEU

Name this task. Dashes and Underscores are ok. No Spaces

Specify the Language Pair

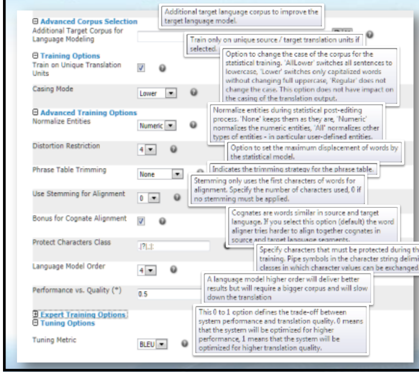
Select the 3 files: training, testing and tuning subsets of Corpus. Or…

Faster: Select just once the directory which contains the 3 files. Done.

Options, Advanced Options, and Expert options, as well as Tuning options!

When ready, submit the new task for processing by the Training server.

Advanced Options



⊟ Advanced Corpus Selection

Additional Target Corpus for Language Modeling

Additional target language corpus to improve the target language model.

⊟ Training Options

Train on Unique Translation Units ☑

Train only on unique source / target translation units if selected.

Casing Mode: Lower

Option to change the case of the corpus for the statistical training. 'AllLower' switches all sentences to lowercase, 'Lower' switches only capitalized words without changing full uppercase, 'Regular' does not change the case. This option does not have impact on the casing of the translation output.

⊟ Advanced Training Options

Normalize Entities: Numeric

Normalize entities during statistical post-editing process. 'None' keeps them as they are, 'Numeric' normalizes the numeric entities, 'All' normalizes other types of entities - in particular user-defined entities.

Distortion Restriction: 4

Option to set the maximum displacement of words by the statistical model.

Phrase Table Trimming: None

Indicates the trimming strategy for the phrase table.

Use Stemming for Alignment: 0

Stemming only uses the first characters of words for alignment. Specify the number of characters used, 0 if no stemming must be applied.

Bonus for Cognate Alignment ☑

Cognates are words similar in source and target language. If you select this option (default) the word aligner tries harder to align together cognates in source and target language segments.

Protect Characters Class: .|?|.:|:

Specify characters that must be protected during the training. Pipe symbols in the character string delimit classes in which character values can be exchanged.

Language Model Order: 4

A language model higher order will deliver better results but will require a bigger corpus and will slow down the translation

Performance vs. Quality (*): 0.5

This 0 to 1 option defines the trade-off between system performance and translation quality. 0 means that the system will be optimized for higher performance, 1 means that the system will be optimized for higher translation quality.

⊞ Expert Training Options

⊟ Tuning Options

Tuning Metric: BLEU

Casing mode: Try AllLower if small Corpus, compare with 'Lower'

Distortion Restriction: 2-4 in many cases, 4-6 for Russian, German, Japanese

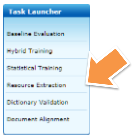
Language Model Order: 4-5, 6 if large corpus

Performance vs. Quality: 1.0

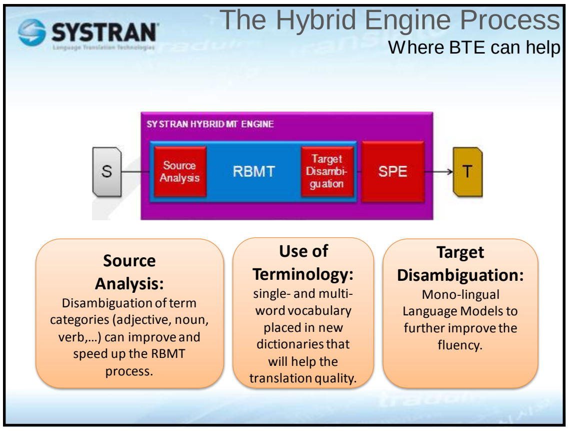
Bilingual Terminology Extraction (BTE)

Automatic Creation of Dictionaries

SES 7 Training Server



Resource Extraction



The Hybrid Engine Process

Where BTE can help

SYSTRAN HYBRID MT ENGINE

S

Source Analysis

RBMT

Target Disambi-guation

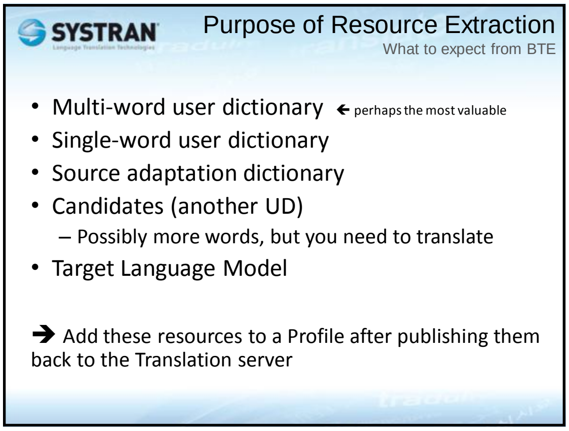
SPE

T

Source Analysis: Disambiguation of term categories (adjective, noun, verb,…) can improve and speed up the RBMT process.

Use of Terminology: single- and multi-word vocabulary placed in new dictionaries that will help the translation quality.

Target Disambiguation: Mono-lingual Language Models to further improve the fluency.

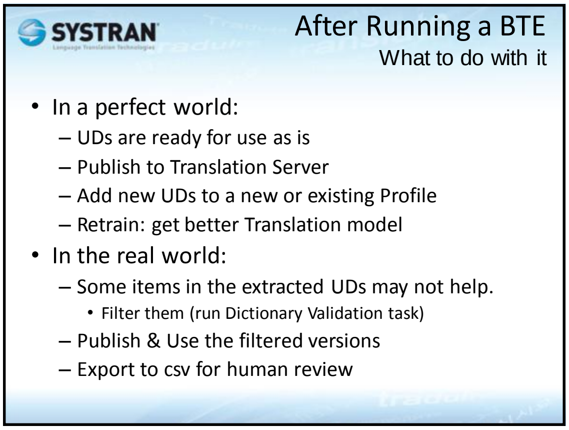


Purpose of Resource Extraction

What to expect from BTE

- Multi-word user dictionary ← perhaps the most valuable
- Single-word user dictionary
- Source adaptation dictionary
- Candidates (another UD)
  - Possibly more words, but you need to translate
- Target Language Model

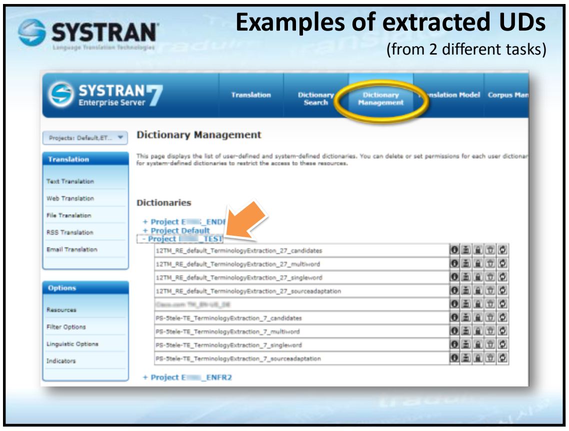
➔ Add these resources to a Profile after publishing them back to the Translation server



After Running a BTE

What to do with it

- In a perfect world:
  - UDs are ready for use as is
  - Publish to Translation Server
  - Add new UDs to a new or existing Profile
  - Retrain: get better Translation model
- In the real world:
  - Some items in the extracted UDs may not help.
    - Filter them (run Dictionary Validation task)
  - Publish & Use the filtered versions
  - Export to csv for human review



Examples of extracted UDs

(from 2 different tasks)

Dictionary Management

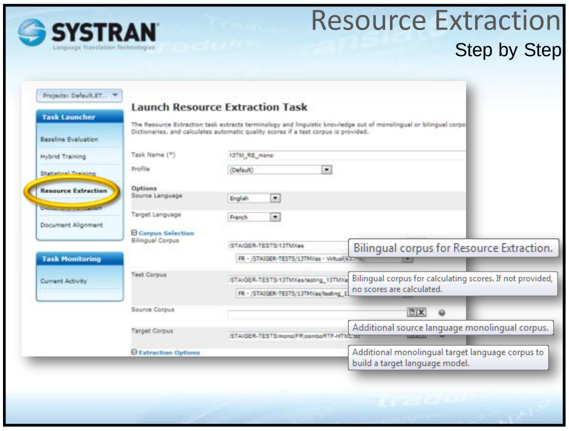
This page displays the list of user-defined and system-defined dictionaries. You can delete or set permissions for each user dictionar for system-defined dictionaries to restrict the access to these resources.

Dictionaries

+ Project E[redacted]_ENDE
+ Project Default
- Project [redacted]_TEST

12TM_RE_default_TerminologyExtraction_27_candidates
12TM_RE_default_TerminologyExtraction_27_multiword
12TM_RE_default_TerminologyExtraction_27_singleword
12TM_RE_default_TerminologyExtraction_27_sourceadaptation
[redacted]
PS-5tele-TE_TerminologyExtraction_7_candidates
PS-5tele-TE_TerminologyExtraction_7_multiword
PS-5tele-TE_TerminologyExtraction_7_singleword
PS-5tele-TE_TerminologyExtraction_7_sourceadaptation

+ Project E[redacted]_ENFR2



Resource Extraction

Step by Step

Launch Resource Extraction Task

The Resource Extraction task extracts terminology and linguistic knowledge out of monolingual or bilingual corpo Dictionaries, and calculates automatic quality scores if a test corpus is provided.

Task Name (*): 12TM_RE_mono
Profile: (Default)

Options
Source Language: English
Target Language: French

⊟ Corpus Selection
Bilingual Corpus: /STAIGER-TESTS/13TMXes
FR - /STAIGER-TESTS/13TMXes - Virtual(63.7%)
Test Corpus: /STAIGER-TESTS/13TMXes/testing_13TMXe
FR - /STAIGER-TESTS/13TMXes/testing_1
Source Corpus:
Target Corpus: /STAIGER-TESTS/mono/FR/comboRTF-HTML.txt

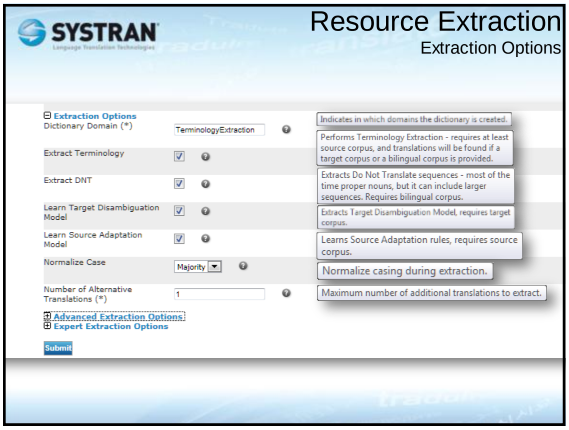
⊟ Extraction Options

Bilingual corpus for Resource Extraction.

Bilingual corpus for calculating scores. If not provided, no scores are calculated.

Additional source language monolingual corpus.

Additional monolingual target language corpus to build a target language model.



Resource Extraction

Extraction Options

⊟ Extraction Options

Dictionary Domain (*): TerminologyExtraction — Indicates in which domains the dictionary is created.

Extract Terminology ☑ — Performs Terminology Extraction - requires at least source corpus, and translations will be found if a target corpus or a bilingual corpus is provided.

Extract DNT ☑ — Extracts Do Not Translate sequences - most of the time proper nouns, but it can include larger sequences. Requires bilingual corpus.

Learn Target Disambiguation Model ☑ — Extracts Target Disambiguation Model, requires target corpus.

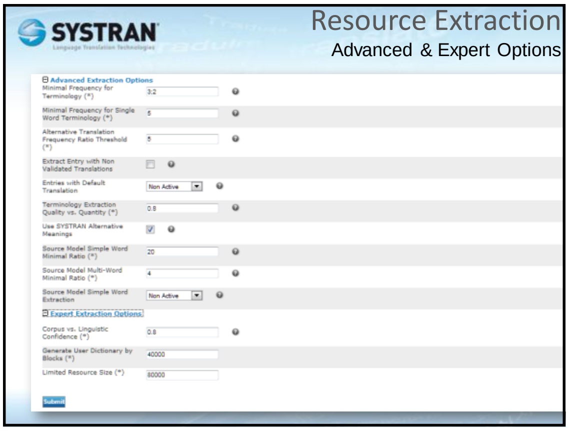
Learn Source Adaptation Model ☑ — Learns Source Adaptation rules, requires source corpus.

Normalize Case: Majority — Normalize casing during extraction.

Number of Alternative Translations (*): 1 — Maximum number of additional translations to extract.

⊞ Advanced Extraction Options
⊞ Expert Extraction Options

Submit



Resource Extraction

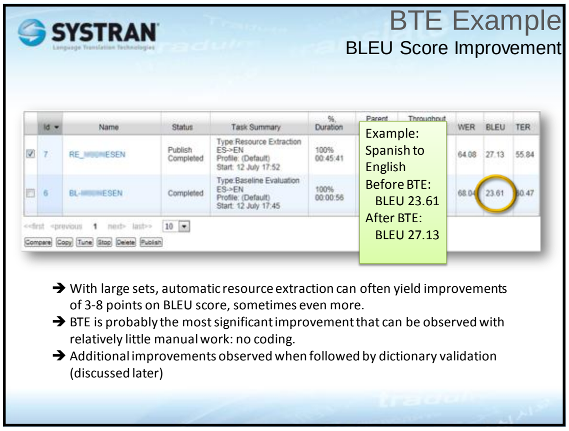
Advanced & Expert Options

⊟ Advanced Extraction Options

Minimal Frequency for Terminology (*)	3:2	?
Minimal Frequency for Single Word Terminology (*)	5	?
Alternative Translation Frequency Ratio Threshold (*)	5	?
Extract Entry with Non Validated Translations	☐	?
Entries with Default Translation	Non Active ▾	?
Terminology Extraction Quality vs. Quantity (*)	0.8	?
Use SYSTRAN Alternative Meanings	☑	?
Source Model Simple Word Minimal Ratio (*)	20	?
Source Model Multi-Word Minimal Ratio (*)	4	?
Source Model Simple Word Extraction	Non Active ▾	?

⊟ Expert Extraction Options

Corpus vs. Linguistic Confidence (*)	0.8	?
Generate User Dictionary by Blocks (*)	40000	
Limited Resource Size (*)	80000	


SYSTRAN
Language Translation Technologies

BTE Example

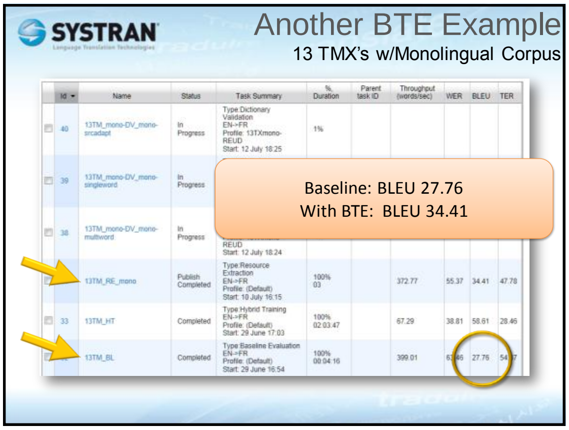
BLEU Score Improvement

	Id ▾	Name	Status	Task Summary	%, Duration	Parent	Throughput	WER	BLEU	TER
☑	7	RE_[redacted]ESEN	Publish Completed	Type:Resource Extraction ES->EN Profile: (Default) Start: 12 July 17:52	100% 00:45:41			64.08	27.13	55.84
☐	6	BL-[redacted]ESEN	Completed	Type:Baseline Evaluation ES->EN Profile: (Default) Start: 12 July 17:45	100% 00:00:56			68.04	23.61	[illegible]0.47

<<first <previous 1 next> last>> 10 ▾
Compare Copy Tune Stop Delete Publish

Example:
Spanish to English
Before BTE:
BLEU 23.61
After BTE:
BLEU 27.13

- ➔ With large sets, automatic resource extraction can often yield improvements of 3-8 points on BLEU score, sometimes even more.
- ➔ BTE is probably the most significant improvement that can be observed with relatively little manual work: no coding.
- ➔ Additional improvements observed when followed by dictionary validation (discussed later)

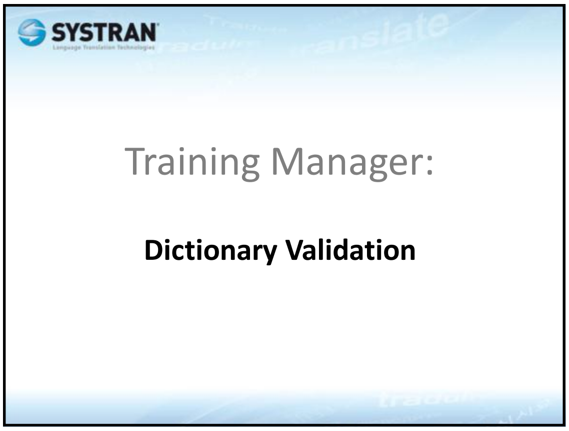


Another BTE Example

13 TMX's w/Monolingual Corpus

	Id ▾	Name	Status	Task Summary	%, Duration	Parent task ID	Throughput (words/sec)	WER	BLEU	TER
☐	40	13TM_mono-DV_mono-srcadapt	In Progress	Type:Dictionary Validation EN->FR Profile: 13TXmono-REUD Start: 12 July 18:25	1%					
☐	39	13TM_mono-DV_mono-singleword	In Progress							
☐	38	13TM_mono-DV_mono-multiword	In Progress	REUD Start: 12 July 18:24						
☐		13TM_RE_mono	Publish Completed	Type:Resource Extraction EN->FR Profile: (Default) Start: 10 July 16:15	100% 03		372.77	55.37	34.41	47.78
☐	33	13TM_HT	Completed	Type:Hybrid Training EN->FR Profile: (Default) Start: 29 June 17:03	100% 02:03:47		67.29	38.81	58.61	28.46
☐		13TM_BL	Completed	Type:Baseline Evaluation EN->FR Profile: (Default) Start: 29 June 16:54	100% 00:04:16		399.01	63.46	27.76	54.[illegible]7

Baseline: BLEU 27.76
With BTE: BLEU 34.41

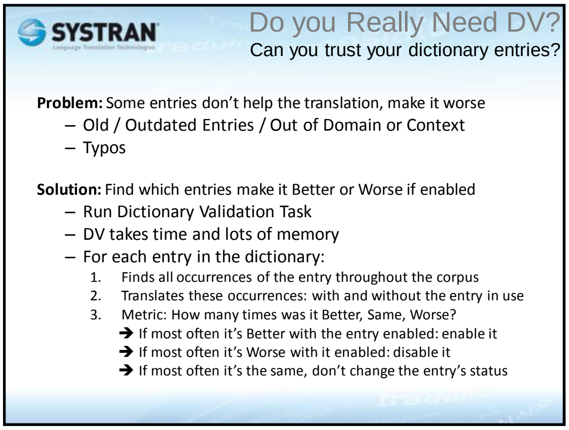


Training Manager:

Dictionary Validation

SES 7
Training Server

Dictionary Validation

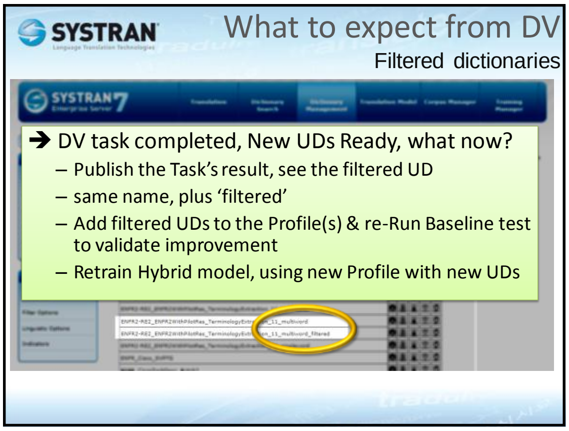


Do you Really Need DV?

Can you trust your dictionary entries?

Problem: Some entries don't help the translation, make it worse
- Old / Outdated Entries / Out of Domain or Context
- Typos

Solution: Find which entries make it Better or Worse if enabled
- Run Dictionary Validation Task
- DV takes time and lots of memory
- For each entry in the dictionary:
  1. Finds all occurrences of the entry throughout the corpus
  2. Translates these occurrences: with and without the entry in use
  3. Metric: How many times was it Better, Same, Worse?
     - → If most often it's Better with the entry enabled: enable it
     - → If most often it's Worse with it enabled: disable it
     - → If most often it's the same, don't change the entry's status



What to expect from DV

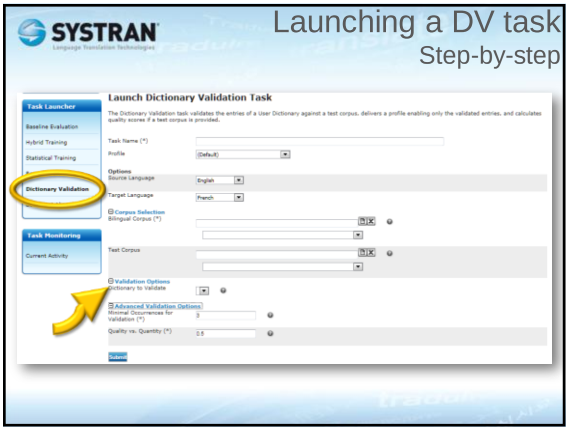
Filtered dictionaries

➔ DV task completed, New UDs Ready, what now?

- Publish the Task's result, see the filtered UD
- same name, plus 'filtered'
- Add filtered UDs to the Profile(s) & re-Run Baseline test to validate improvement
- Retrain Hybrid model, using new Profile with new UDs

ENFR2-RE2_ENFR2WithPilotRes_TerminologyExtr…n_11_multiword

ENFR2-RE2_ENFR2WithPilotRes_TerminologyExt…n_11_multiword_filtered



Launching a DV task

Step-by-step

Launch Dictionary Validation Task

The Dictionary Validation task validates the entries of a User Dictionary against a test corpus, delivers a profile enabling only the validated entries, and calculates quality scores if a test corpus is provided.

Task Name (*)

Profile: (Default)

Options

Source Language: English

Target Language: French

⊟ Corpus Selection

Bilingual Corpus (*)

Test Corpus

⊟ Validation Options

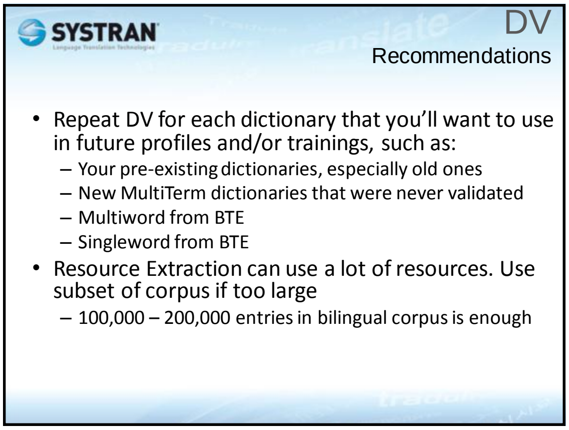
Dictionary to Validate

⊟ Advanced Validation Options

Minimal Occurrences for Validation (*): 3

Quality vs. Quantity (*): 0.5

SYSTRAN
Language Translation Technologies
Task Launcher
Baseline Evaluation
Hybrid Training
Statistical Training
Dictionary Validation
Task Monitoring
Current Activity
Submit



DV

Recommendations

- Repeat DV for each dictionary that you'll want to use in future profiles and/or trainings, such as:
  - Your pre-existing dictionaries, especially old ones
  - New MultiTerm dictionaries that were never validated
  - Multiword from BTE
  - Singleword from BTE
- Resource Extraction can use a lot of resources. Use subset of corpus if too large
  - 100,000 – 200,000 entries in bilingual corpus is enough

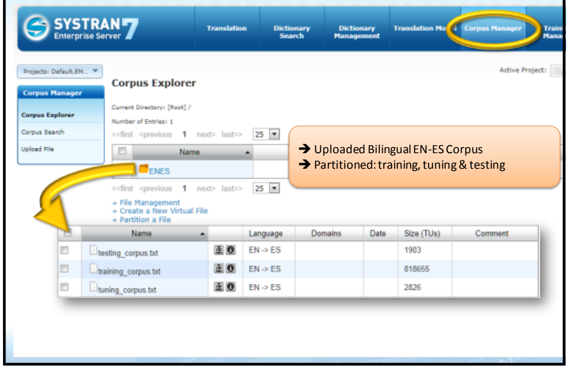
SYSTRAN
Language Translation Technologies

Putting it all together

SYSTRAN
Language Translation Technologies

Complete Cycle (1/4)

Bilingual Corpus, English-Spanish



Corpus Explorer

Current Directory: [Root] /

Number of Entries: 1

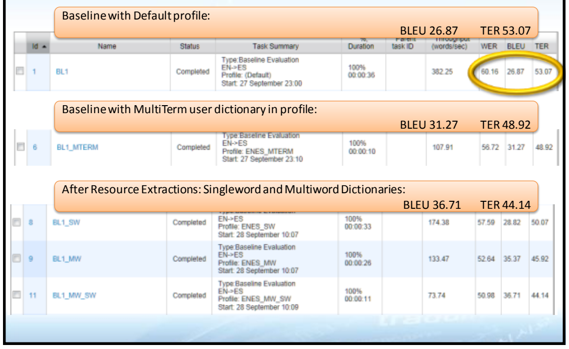
➔ Uploaded Bilingual EN-ES Corpus
➔ Partitioned: training, tuning & testing

□	Name		Language	Domains	Date	Size (TUs)	Comment
□	testing_corpus.txt		EN -> ES			1903	
□	training_corpus.txt		EN -> ES			818655	
□	tuning_corpus.txt		EN -> ES			2826	

SYSTRAN
Language Translation Technologies

Complete Cycle (2/4)

Baseline tests & Resource Extraction



Baseline with Default profile:
BLEU 26.87 TER 53.07

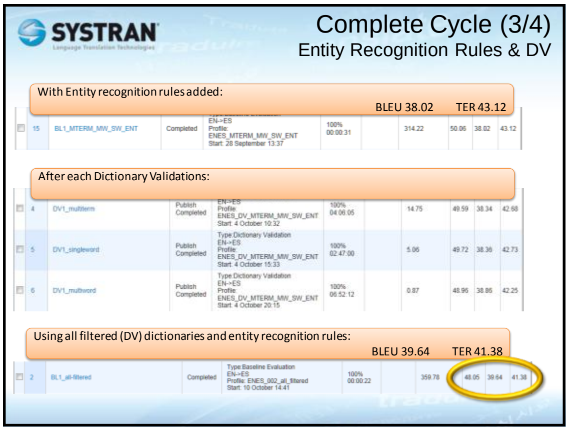
□	Id	Name	Status	Task Summary	%, Duration	Parent task ID	Throughput (words/sec)	WER	BLEU	TER
□	1	BL1	Completed	Type:Baseline Evaluation EN->ES Profile: (Default) Start: 27 September 23:00	100% 00:00:36		382.25	60.16	26.87	53.07

Baseline with MultiTerm user dictionary in profile:
BLEU 31.27 TER 48.92

□	Id	Name	Status	Task Summary	%, Duration	Parent task ID	Throughput (words/sec)	WER	BLEU	TER
□	6	BL1_MTERM	Completed	Type:Baseline Evaluation EN->ES Profile: ENES_MTERM Start: 27 September 23:10	100% 00:00:10		107.91	56.72	31.27	48.92

After Resource Extractions: Singleword and Multiword Dictionaries:
BLEU 36.71 TER 44.14

□	Id	Name	Status	Task Summary	%, Duration	Parent task ID	Throughput (words/sec)	WER	BLEU	TER
□	8	BL1_SW	Completed	EN->ES Profile: ENES_SW Start: 28 September 10:07	100% 00:00:33		174.38	57.59	28.82	50.07
□	9	BL1_MW	Completed	Type:Baseline Evaluation EN->ES Profile: ENES_MW Start: 28 September 10:07	100% 00:00:26		133.47	52.64	35.37	45.92
□	11	BL1_MW_SW	Completed	Type:Baseline Evaluation EN->ES Profile: ENES_MW_SW Start: 28 September 10:09	100% 00:00:11		73.74	50.98	36.71	44.14



Complete Cycle (3/4)

Entity Recognition Rules & DV

With Entity recognition rules added:
BLEU 38.02 TER 43.12

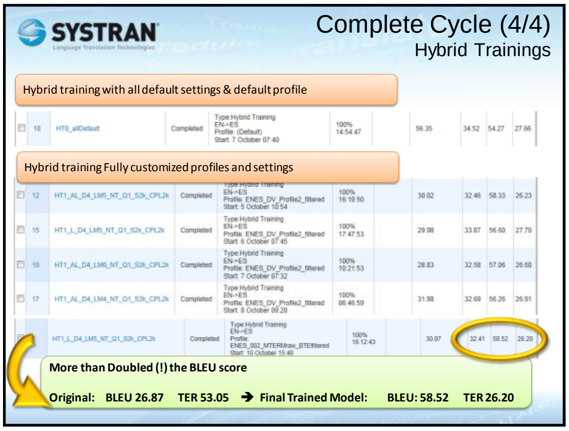
□	15	BL1_MTERM_MW_SW_ENT	Completed	EN->ES Profile: ENES_MTERM_MW_SW_ENT Start: 28 September 13:37	100% 00:00:31		314.22	50.06	38.02	43.12

After each Dictionary Validations:

□	4	DV1_multiterm	Publish Completed	EN->ES Profile: ENES_DV_MTERM_MW_SW_ENT Start: 4 October 10:32	100% 04:06:05		14.75	49.59	38.34	42.68
□	5	DV1_singleword	Publish Completed	Type:Dictionary Validation EN->ES Profile: ENES_DV_MTERM_MW_SW_ENT Start: 4 October 15:33	100% 02:47:00		5.06	49.72	38.36	42.73
□	6	DV1_multiword	Publish Completed	Type:Dictionary Validation EN->ES Profile: ENES_DV_MTERM_MW_SW_ENT Start: 4 October 20:15	100% 06:52:12		0.87	48.96	38.86	42.25

Using all filtered (DV) dictionaries and entity recognition rules:
BLEU 39.64 TER 41.38

□	2	BL1_all-filtered	Completed	Type:Baseline Evaluation EN->ES Profile: ENES_002_all_filtered Start: 10 October 14:41	100% 00:00:22		359.78	48.05	39.64	41.38



Complete Cycle (4/4)

Hybrid Trainings

Hybrid training with all default settings & default profile

□	18	HT0_allDefault	Completed	Type:Hybrid Training EN->ES Profile: (Default) Start: 7 October 07:40	100% 14:54:47		56.35	34.52	54.27	27.66

Hybrid training Fully customized profiles and settings

□	12	HT1_AL_D4_LM5_NT_Q1_S2k_CPL2k	Completed	Type:Hybrid Training EN->ES Profile: ENES_DV_Profile2_filtered Start: 5 October 10:54	100% 16:19:50		30.02	32.46	58.33	26.23
□	15	HT1_L_D4_LM5_NT_Q1_S2k_CPL2k	Completed	Type:Hybrid Training EN->ES Profile: ENES_DV_Profile2_filtered Start: 6 October 07:45	100% 17:47:53		29.98	33.87	56.60	27.70
□	16	HT1_AL_D4_LM6_NT_Q1_S2k_CPL2k	Completed	Type:Hybrid Training EN->ES Profile: ENES_DV_Profile2_filtered Start: 7 October 07:32	100% 10:21:53		28.83	32.58	57.06	26.60
□	17	HT1_AL_D4_LM4_NT_Q1_S2k_CPL2k	Completed	Type:Hybrid Training EN->ES Profile: ENES_DV_Profile2_filtered Start: 8 October 09:28	100% 06:46:59		31.98	32.69	56.26	26.91
□	2	HT1_L_D4_LM5_NT_Q1_S2k_CPL2k	Completed	Type:Hybrid Training EN->ES Profile: ENES_002_MTERMraw_BTEfiltered Start: 10 October 15:49	100% 16:12:43		30.07	32.41	58.52	26.20

More than Doubled (!) the BLEU score

Original: BLEU 26.87 TER 53.05 ➔ Final Trained Model: BLEU: 58.52 TER 26.20

Conclusion

This completes our exploration of how to increase localization efficiency with SYSTRAN Hybrid MT Products.
Here are some of the topics we've covered:

- Translation Profiles
- User Dictionaries (UD)
- Normalization Dictionaries (ND)
- Translation Memories (TM)
- Linguistic Resources
- Resource Extractions (BTE)
- Dictionary Validation (DV)
- Hybrid Training, Statistical Training