Dictionary and Glossary Creation Guidelines
Version 7.0

Contents

1 Introduction

Key to improving translation accuracy (whether machine or human translation) is in identifying, translating and publishing dictionaries (term bases) which are used by Translution's software. The same dictionaries can also be used to improve the quality of human translation.

2 Types of Dictionary Data

There are three different types of dictionary data which are used by Translution's software.

- Do Not Translate (DNTs): Terms which should not be translated.

  DNT's typically consist of contacts, customers, suppliers, addresses, product names, brand names etc

- Terms: Terms which are specific to your industry and company.

  Terms are normally derived from mono-lingual electronic data and are then translated by terminologists to create bi-lingual dictionaries for each language pair.

  Terms are defined with an attribute of the part of speech of either nouns, proper nouns, compound terms, acronyms, abbreviations or phrases

- Translation Memory (TM): Translation Memory (TM) is source and target human translations stored and aligned into segments (usually sentences or phrases).

  Using TM for machine translation can deliver consistent and perfect translations of standard phrases used by your organization.

  Translation Memory is also created by translators when using Translution Localization Manager.

  Storing, managing and using Translation Memory reduces the cost of translation services because source language matches do not need to be re-translated. It also ensures your translations are consistent.

3 Dictionary Types

Translution provide 3 different types of dictionaries.

Glossary
This type of dictionary contains DNTs (obtained from the customers source documentation) and any meta data (frequently used terms or keywords) derived from the customer website pages or as supplied by the customer which require translation.

Note that a glossary is not sufficient to enhance machine translation quality and is therefore not suitable where post edited translation is required or when using Translution Business, Translution DB Manager (with machine translation) or Translution MT API.

A glossary will however ensure that a customer’s key terms are consistently translated by translators.

Dictionary

A dictionary includes a glossary. However it is usually larger and therefore can be used to improve machine translation quality.

A dictionary is derived from customers documents and includes the most frequently used terms and phrases..

We recommend that you invest in creating a dictionary when one or more of the following applies.

- You have licensed Translution Business or Translution API and wish to improve the quality of the machine translation
- You have licensed Translution Localization Manager for use by your own teams of translators and wish to improve the productivity of your translators.
- You have a requirement for large amounts of translation over a period (typically in excess of 50,000 source words).

Dictionaries are normally created from English source material. For other source languages please enquire.

Note that dictionaries and glossaries may be reversed i.e. you can create a French to English dictionary from an English to French dictionary.

A 1000 term dictionary should normally be sufficient to improve machine translation quality. A bigger dictionary is normally recommended for technical translation or for large websites. Translution can recommend the size of dictionary required after reviewing your requirements.

Translation Memory

Some customers alreadyy have a translation memory. Provided this is in standard TMX format, it usually can be published by Translution and ensures an immediate improvement in bith human and machine translation quality.

Alternatively we can align existing bilingual data can be aligned to create a Translation Memory.

4 Process

Translution offer a structured approach to creating your dictionaries.

This consists of the following steps:

Stage 1. Documentation Collection

We first collect electronic data from the customer.

Potential sources are:

- Existing Departmental Documents (Word, Excel format)
- Web Site/Intranet (HTML format)

- Brochures (Word or PDF format)
- Technical/Quality Manuals (Word format)
- Quotations and Proposals (Word format)
- Price Lists (Excel)
- Contacts (Excel) – structured please
- Custom Dictionaries (Excel)
- Database extracts (Excel)
- Any Bilingual data already available (TMX only please if provided by another translation agency)

Glossaries (plus some DNT's) should be provided as spreadsheets of terms. Spreadsheets should be first checked by the customer and should include terms that require translation and spreadsheets of Contacts, Customers, Suppliers, Addresses, and/or Product Names.

For small dictionaries (less than or equal to 1000 terms) please follow these guidelines:

- Create a single text file containing text data from multiple sources to encapsulate the company's main domain areas (e.g. Existing Corporate Documents, Brochures, Specifications, Training Material, Procedures, etc.)
- Web Site/Intranet Files (HTML format) [**not lists of URLs**].
- 1 or 2 spreadsheets* of DNTs (Contacts, Customers, Suppliers, Addresses, and/or Product Names) either from existing spreadsheet data or a database.

For larger dictionaries (more than 1000 terms)

- Documents from multiple sources can b accepted which encapsulate the company's main domain areas (e.g. Departmental documents, existing corporate documents, Brochures, Specifications, Training Material, Procedures, etc.) [Plain Text, Word format, **no PDFs**]
- Web Site/Intranet Files (HTML format) [**no list of URLs**].
- Up to 10 spreadsheets of DNTs (Contacts, Customers, Suppliers, Addresses, and/or Product Names) either from existing spreadsheet data or extracted from a database.
- Any available bilingual data in a structured format may be included.

Note: A spreadsheet is assumed to contain a single sheet of data.

The documents are normally transfered using FTP to Translution or provided on CD by the client.

In all cases please organise your data into logical folders which are clearly marked.

Stage 2. Conversion into Text format

We then convert all documents and data into text format to be used by Translution's term extraction software.

Stage 3. Merging Text Documents

We then merges all the text into a single plain text file

Stage 4. Term Extraction

We then extract the base form of terms (nouns, proper nouns, noun phrases and acronyms) from the merged text file using Translution's term extraction software. This is a special program developed by Translution's computational linguists which:

- Generates automatically the base form of all nouns, proper nouns and noun phrases
- Counts the frequency of all terms extracted from the data.
- Excludes the most commonly used nouns in English. This database is a unique resource developed by Translution which typically reduces the database of terms by 40%, leaving the most likely candidates as possible terms.

A spreadsheet is then reviewed and classified by us to produce a suitable spreadsheet for customer reviewing.

Stage 5. Spreadsheet Review by Customer

You personnel review the extracted term spreadsheet and mark them as "Translate", "Do Not Translate" and "Ignore"

Terms are split into separate worksheets for nouns, proper nouns, noun phrases, contacts and any existing defined terms and their frequencies.

Terms identified as Translate may require you to, provide a definition where they have a specific context.

Terms identified as Do Not Translate ideally also require a definition e.g. Contact, Company Name etc.

No further information is required for terms that are marked as Ignore.

The completed spreadsheet is then signed off and returned to Translution.

Stage 6. Translation of Agreed Terms

Translution translates the source terms using its MT systems and any industry specialised dictionaries (such as its Business term dictionaries). Translution sends the terms marked as Translate together with definitions and any translation to a specialist terminologist, who translates them and verifies translations and returns them to Translution.

Stage 7. Publishing of your Dictionary

Translate terms and Do Not Translate terms are then uploaded and publsihed for use by Translution Localization Manager and compiled for use by our machine translation engines.

The completed dictionary is also sent to you as an excel spreadsheet.