Crisis Translation

Lesson 11: Machine Translation & Corpora Management

Dr Emmanouela Patiniotaki

HELLENIC REPUBLIC
National and Kapodistrian University of Athens

About the module

Lesson	Thematic Units	
1	Introduction to Crisis and Crisis Translation	Understanding crisis
2	Crisis Policies & Communication	Understanding stakeholders
3	Language and Translation as a means of communication in Crisis	Understanding language
4	Ethics in Crisis Translation	Project Management in Crisis
5	Greek Crisis Management and Policies	Controlled Language
6	Interpreting and Translation in Crisis	Training resources
7	Translators with or without resources in Crisis	Pre-editing for MT
8	Translation stages in Crisis – Preparedness	Post-editing for MT
9	Access to political and social resources	Translating for Immigration
10	The role of Social Media	Translating for Disability
11	Machine Translation Quality	Translating in Medical contexts
12	Speed and collaboration	Translating Guidelines
	Practical translation topics will be spread within various units	©2021 Patiniotaki

Previous Task

Mock Translation – Short & simple versions – Social Media posts

Scenario

You work for INFOCRISIS.GOV.

An urgent announcement has been sent and it needs to be distributed on the web and in social media within the next hour.

You have 523 words in the source.

Πυροσβεστικό Σώμα: Σε επιχειρησιακή ετοιμότητα λόγω πρόβλεψης Πολύ Υψηλού Κινδύνου Πυρκαγιάς για τις 31/8 και ενημέρωση για τις πυρκαγιές της 30ης/8

30 Αυγούστου 2019

Share

Task

You need to prepare:

1. A summarized translation (200-250 words)
2. A short **guide** for citizens in Greek
3. A short **guide** for citizens in English
4. A **post** for social media in both (up to 280 characters)

Timing

You need to prepare:

1. A summarized translation
2. A short **guide** for citizens EL
3. A short **guide** for citizens EN
4. A **post** for social media
   (up to 280 characters)

Crisis Translation Review

- Make notes on the tools you used
- Make notes on the time each task took
- Makes notes on the main difficulties you came across

Machine-translated

Το 2ο στάδιο επιχειρησιακής ετοιμότητας των δυνάμεών του θέτει σε εφαρμογή το Πυροσβεστικό Σώμα για αύριο Σάββατο 31 Αυγούστου 2019 για τις περιοχές στις οποίες προβλέπεται Πολύ Υψηλός Κίνδυνος Πυρκαγιάς (κατηγορία κινδύνου 4) σύμφωνα με τον Χάρτη Πρόβλεψης Κινδύνου Πυρκαγιάς που εξέδωσε νωρίτερα σήμερα η Γενική Γραμματεία Πολιτικής Προστασίας.

Στις περιοχές αυτές τίθεται επίσης σε εφαρμογή και το μέτρο της προληπτικής απαγόρευσης της κυκλοφορίας οχημάτων και της παραμονής εκδρομέων σε εθνικούς δρυμούς, δάση και «ευπαθείς» περιοχές, σύμφωνα με το Σχέδιο Δράσης της Πολιτικής Προστασίας για την αντιμετώπιση κινδύνων λόγω δασικών πυρκαγιών. Όπως τονίζεται σε σχετική ενημέρωση από το Π.Σ. θα συνεχιστούν παράλληλα οι περιπολίες εναέριας επιτήρησης, καθώς και οι μικτές περιπολίες από Πυροσβεστικές, Αστυνομικές και Στρατιωτικές δυνάμεις.

The 2nd stage of operational readiness of its forces is implemented by the Fire Brigade for tomorrow Saturday 31 August 2019 for the areas in which a Very High Fire Risk is foreseen (risk category 4) according to the Fire Risk Forecast Map issued by the Fire Brigade today. Protection.

In these areas, the measure of preventive prohibition of vehicle traffic and the stay of hikers in national parks, forests and "vulnerable" areas is also implemented, according to the Action Plan of the Civil Protection to deal with risks due to forest fires. As emphasized in a relevant briefing by the PS. Aerial surveillance patrols will continue at the same time, as well as joint patrols by the Fire, Police and Military Forces.

What is Machine Translation?

- Fields involved:
  - Lexicography
  - Linguistics
  - Computational Linguistics
  - Computer Science
  - Language Engineering

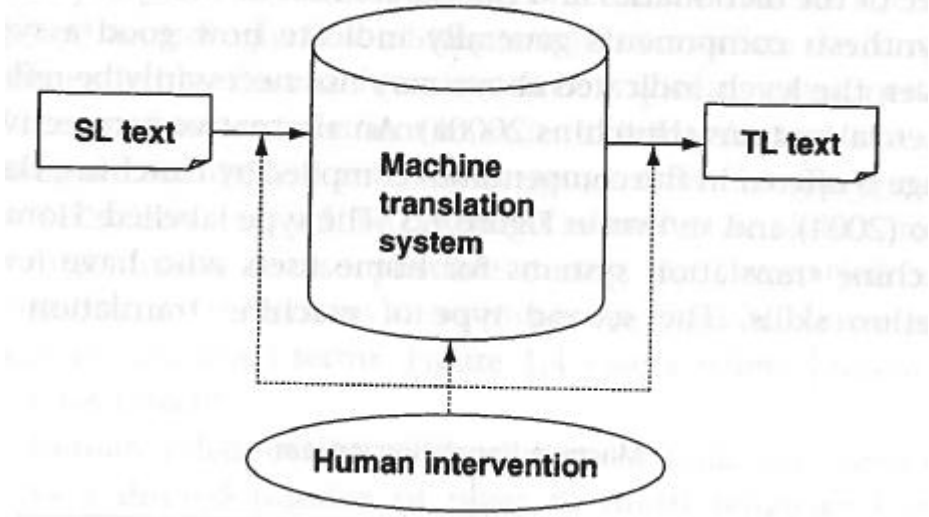
What is Machine Translation?

- “the application of computers to the task of translating texts from one natural language to another” (EAMT)
- “the attempt to automate all *or part of* the process of translating from one human language to another” (Arnold et al., 1994: 1)
- “computerised systems responsible for the production of translations from one natural language into another, with or without human assistance” (Hutchins & Somers, 1992: 3)

What is Machine Translation?

- MT is based on the hypothesis that natural languages can be fully described, controlled and mathematically coded. (Wills, 1999: 140)

The issue of human intervention:



Adopted from Quah (2006: 9)

Other useful definitions

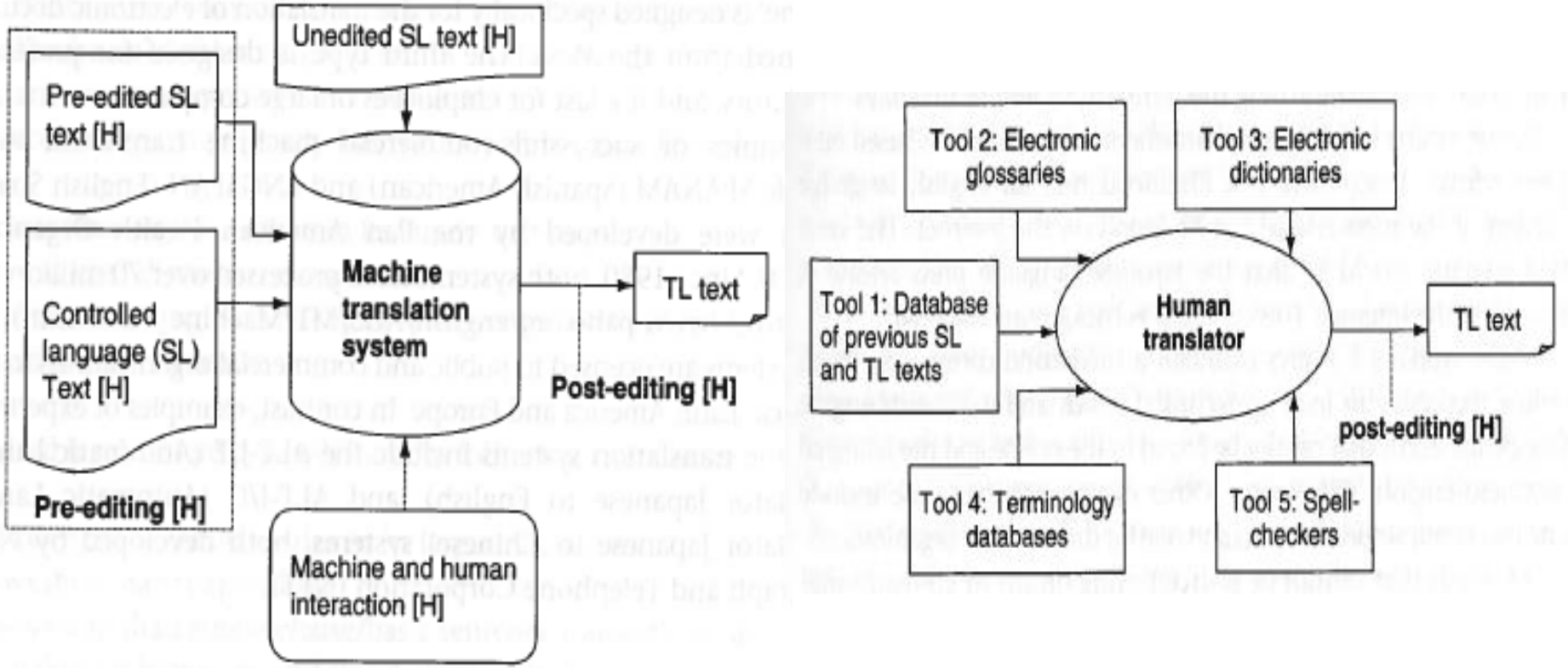
- Pre-editing: identification of elements, e.g. idioms and typos, that may cause problems for the MT system during the translation process
- Post-editing: correction of the output of MT in order to achieve the required language standards (including style, terminology, format, etc.)
- Controlled language: follows strict grammar rules and vocabulary
- H: human SL: Source Language TL: Target Language

Human-aided VS Machine-aided

- Corpora
- TMs & TBs
- Pre-edited SL
- Controlled SL

- Electronic dictionaries
- Electronic glossaries
- Spell-checkers
- Post-editing

Input VS Output: [H]/[M]?

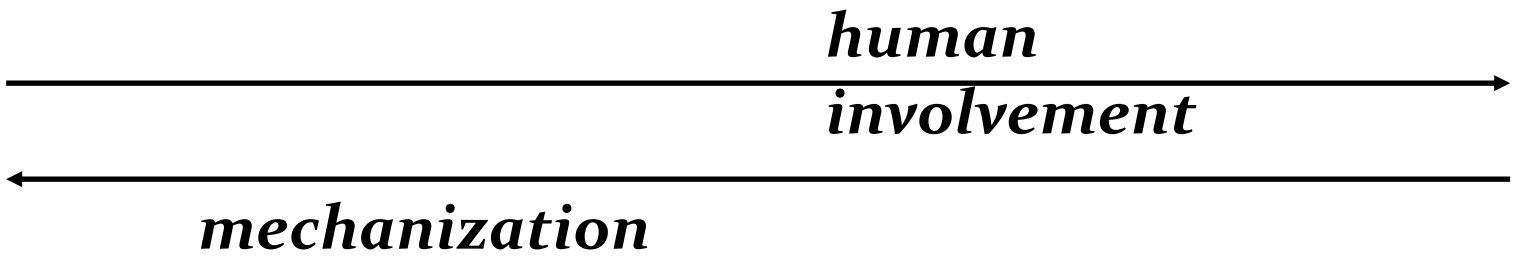


Adopted from Quah (2006: 12-13)

Types of Machine Translation

- **FAMT**: Fully Automated Machine Translation: performed without the intervention of human beings during the process.
- **FAHQMT**: Fully Automated High Quality Machine Translation.
- **CAT**: Computer-aided Translation.
  - **HAMT**: Human-Aided Machine Translation, MT systems that depend on the input of humans.
  - **MAHT**: Machine-aided Human Translation (e.g. TM).

Types of Machine Translation



Fully Automated Translation	Human-Aided Machine Translation	Machine-Aided Human Translation	Traditional Human Translation

Computer-Assisted Translation (CAT)

Adapted from Austermühl (2001: 10)

Basic Facts and Misconceptions

"The quality of translation you can get from an MT system is very low, so it's useless"

"Back translation shows that MT is useless"

The spirit is willing, but the flesh is weak →
The vodka is good, but the steak is lousy

"MT is a waste of time, a machine will never translate Shakespeare"

"MT threatens the jobs of translators"

Source: Arnold et al. 1994

Why MT matters?

Social and Political Importance

- To facilitate communication

Economic importance

- It facilitates commerce

Scientific importance

- Testing ground for AI and linguistic theories

Philosophical importance

- As an attempt to automate an activity that requires the full range of human knowledge

UCL

Source: Arnold et al. 1994

Reasons for using MT

Translation for dissemination

- traditional need for translations of 'publishable' quality.

Translation for assimilation

- demand for translation of short-lived documents for information gathering and analysis.

Translation for interpersonal communication

- demand for on-the-spot translation, traditionally met by interpreters, but now required in different environments (e.g. online).

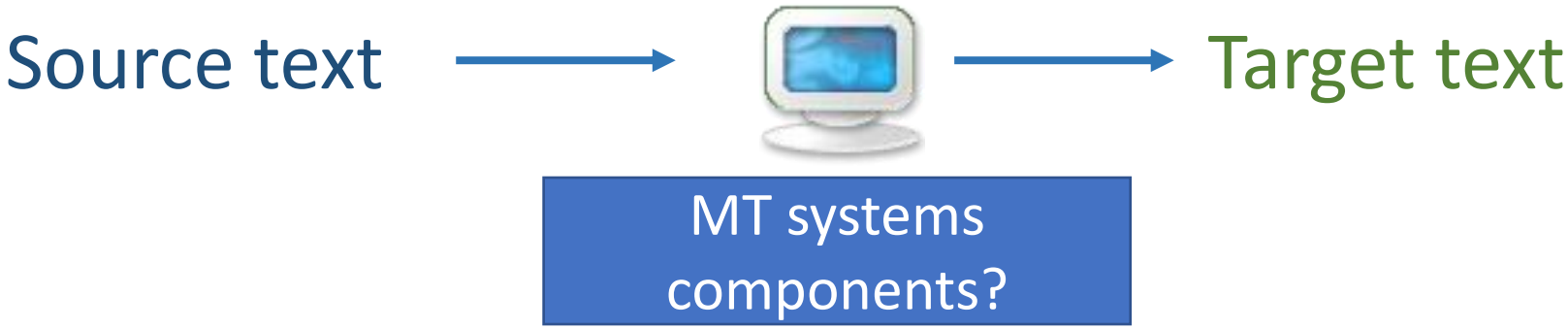
Access to information

- machine translation can facilitate information search and retrieval from textual or non-textual databases.

UCL

Source: Quah, 2006: 89-90

How do MT systems work?



1. **Set of monolingual and bilingual dictionaries**
   - Monolingual dictionaries present grammatical information (morphology, syntax and semantics).
   - Bilingual dictionaries are consulted when a SL word is subsequently matched to its TL equivalent.
2. **Parsers:** a parser assigns a structure to each string in the SL text based on the stored grammatical information already pre-determined for that language. The goal of the parser is to identify the relationships between SL words and their structural representations.

Source: Quah, 2006: 67

References

- Arnold, D., Balkan, L., Humphreys, R. L., Meijer, S. and L. Sadler (1994) *Machine Translation: An Introductory Guide*. Manchester & Oxford: NCC Blackwell
- Austermühl, F. (2001) *Electronic Tools for Translators*. Manchester: St Jerome
- European Association for Machine Translation. http://www.eamt.org/mt.php
- Hutchins, J. W., Somers, H. (1992) An Introduction to Machine Translation. London: Academic Press Inc.
- Koehn & Knowles (2017) *Six challenges for Neural Machine Translation*. https://arxiv.org/pdf/1706.03872.pdf
- Nirenburg, S., Somers, H., Wilks, Y. (eds.) (2003) *Readings in Machine Translation.* Massachusetts: MIT Press.
- Quah, C.K. (2006), *Translation and Technology*. Palgrave: Macmillan.
- Somers, H. (ed.) (2003) *Computers and Translation: A translator's guide*. Amsterdam/Philadelphia: John Benjamins
- Trujilo, A. (1999) *Translation Engines: Techniques for Machine Translation*. London: Springer
- Wilks, Y. (2009) *Machine translation: its scope and limits*. New York: Springer

What is a corpus?

- A collection of *written* or *spoken* material in machine-readable form, assembled for the purpose of linguistic research. (Oxford Dictionary)
- A collection of *pieces of language text* in electronic form, selected according to *external criteria* to represent, as far as possible, a language or language variety as a source of data for linguistic research. (Sinclair, 2005)

Where & why do we use corpora?

- The most authentic language resource
- For reference in translation (in-context use of language)
- To calculate frequency of elements/structures
- For research purposes (quantitative & qualitative analysis)
- To train language engines
- To facilitate language management tools

Types of corpora

- General (per language) vs specialized corpora (domain or genre specific)
- Monolingual, bilingual and multilingual corpora, parallel corpora
- Comparable corpora (similarity among varieties)
- Spoken and written corpora
- Synchronic and diachronic corpora
- Native speaker and learner corpora

Basic terminology

- ***Token*** – number of all elements in a string
- ***Type*** – number of distinct elements in a string
- ***Lemma*** – an abstract form of a word
- ***Stem*** – the root shared among various forms
- ***Word-form*** – various forms of the same lemma
- ***Utterance*** – spoken string
- ***Sentence*** – written string
- ***N-grams*** – sequence of n items

Examples

1) The cat is black and the dog is white.
   9 *tokens*, 7 *types*

1) Produce, producing, produced
   Stem: produc-
   Lemma: produce
   Word-forms: producing, produced

N-grams

- Items in n-grams can be anything that is specified as a gram, e.g. phonemes, syllables, letters, words etc.

- Example at letter level: use
  …is a 3-gram sequence
- Example at sentence level: He used the pen
  …is a 4-gram sequence

N-grams

- For computers, the easiest way to break a string down into components is to consider substrings
- An n-word substring is called an n-gram.
- If n=2 → bigram
- If n=3 → trigram
- If n=1 → unigram OR word!
- If a string has a lot of reasonable n-grams, then maybe it is a reasonable string.

Task

- Using the Corpus of Contemporary American English and the help info or your handout, find the following:
  - The frequency of the term *calibration*
    (Use the LIST feature)
  - The frequency of the term *work* as a *noun* and as a *verb*
    (Use POS)
  - The frequency of the term *machine translation* in ACADEMIC corpora (Use the CHART feature)
  - The number of occurrences of *white* as a collocate of the word *black* in 3-gram sequences

Pre-editing and post-editing

Pre-editing involves preparing a source text to be translated with a Machine Translation engine in order to avoid problems from the outset. (Austermühl, 2001: 163)

The task of the **post-editor** is to edit, modify and/or correct pre-translated text that has been processed by a MT system from a source language into a target language. (Allen, 2003: 297)

Both tasks are often done by translators
(technical writers are also responsible for pre-editing)

Pre-editing

- Involves applying **controlled language rules** or verifying conformance (controlled language checkers)
- Example of controlled language rules:
  - Avoid idiomatic expressions
  - Avoid omitting pronouns
  - Avoid omitting relative pronouns (e.g. that, who, whom)
  - Keep sentence structures clear, simple and direct
  - Break up long sentences into shorter ones (one idea per sentence)
  - Keep to the typical word order *subject* - *verb* - *object* and avoid the passive voice
  - Avoid splitting separable English verbs (e.g. look up)
  - Keep to standard, formal English in which grammatical connections are clearly expressed
- More rules:
  - http://www.muegge.cc/controlled-language.htm

Post-editing

The task of the post-editor is to edit, modify and/or correct pre-translated text that has been processed by a MT system from a source language into a target language. (Allen, 2003: 297)

Usually post-editing is **done by translators**. This was not initially foreseen by MT pioneers – they thought that anybody could do revision if they knew the target language. But it was quickly realised that revisers do need to know the source language in order to do revision. Post-editing depends largely on bilingual skills acquired over time, and translators have these skills more than non-translators (Hutchins, 2005: 7).

Types of post-editing

Light/Fast/Minimal post-editing

- Quick turn-around
- Essential corrections only

Full post-editing (conventional post-editing)

- Slower turn-around
- More corrections leading to higher quality

General post-editing guidelines

- **Retain as much raw translation as possible**
  - Even if you think writing the translation from scratch will be quicker
- **Don’t hesitate too long over a problem**
- **Don’t worry about style (?)**
- **Don’t embark on time-consuming research**
- **Make changes only where absolutely necessary (essential vs. Preferential changes)**
  - Correct words or phrases that are
    - Nonsensical
    - Wrong
    - Omitted or added unnecessarily
    - If there’s enough time, ambiguous

O’Brien, 2010

References

- Allen, Jeffrey (2003). “Post-editing”. In: Harold Somers (ed.) *Computers and translation: a translator’s guide.* Amsterdam/Philadelphia: John Benjamins , 297-317.
- Allen, Jeffrey (2005). “What is Post-editing?”, in *Translation Automation Newsletter*, Issue 4. February 2005.
- De Almeida, Giselle & O'Brien, Sharon (2010) “Analysing Post-Editing Performance: Correlations with years of translation experience”, *EAMT 2010 - European Association for Machine Translation.*
- Hutchins, John. (2005), “Current commercial machine translation systems and computer-based translation tools: system types and their uses”. *International Journal of Translation* vol.17, no.1-2, Jan-Dec 2005, 5-38.
- Krings, Hans (2001). *Repairing Texts: Empirical Investigations of Machine Translation Post-Editing Processes*. Kent, Ohio: The Kent State University Press.
- O’Brien, Sharon (2010). Introduction to Post Editing, who, what, how and where to next, AMTA 2010, http://amta2010.amtaweb.org/AMTA/papers/6-01-ObrienPostEdit.pdf.
- Vasconcellos, Muriel & Marjorie Léon (1985), “SPANAM and ENGSPAN: Machine Translation at the Pan American Health Organization”, *Computational Linguistics* 11, 122-136