European Language Resource Coordination (ELRC)

Stelios Piperidis
Athena RC/ILSP
Context

- Nobody speaks all languages
- Make sure languages do not create silos or invisible borders
- Restricting mobility of people, information, ideas, commerce
- No language discrimination but equal opportunities
- Languages as an asset, not a problem
• How?
• Translation!
• Extremely skilled profession

• A lot to translate ... in fact much too much
• Technology support
• Human translation supported by technology = Automated Translation (AT)
Machine Translation Evolution

1950-1980
RBMT, KBMT, Interlingua

1984
Example-based MT

1989-1993
Statistical MT (first steps-word-based)

2003-2013
Statistical MT (phrase-based)

2014-2017
Neural MT (first steps)

???
Methodological evolution

**Expert System**
- Experts
- Manually coded rules:
  - If « ... » then ... if « ... » then ...
  - ......
- Expert system output: T: But where are the snows of yesteryear? $P = 0.41$
  - T2: However, where are yesterday’s snows? $P = 0.33$
  - T3: Hey - where did the old snow go? $P = 0.18$
  - ......

**Machine Learning System**
- Bilingual parallel corpus: $S$, $T$
- Learning algorithm
- ML rules:
  - $P (\text{but} | \text{mais}) = 0.7$
  - $P (\text{however} | \text{mais}) = 0.3$
  - $P (\text{where} | \text{où}) = 1.0$
  - ...
<table>
<thead>
<tr>
<th>Rationalism</th>
<th>Empiricism</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Emphasizes on theory</td>
<td>1. Emphasizes on data</td>
</tr>
<tr>
<td>2. Assumes an “innate language faculty”</td>
<td>2. Assumes all knowledge gathered only via senses</td>
</tr>
<tr>
<td>3. Aims at discovering the language of the human mind (linguistic competence)</td>
<td>3. Aims at analysing language as it actually occurs (linguistic performance)</td>
</tr>
<tr>
<td>4. Assigns categories to language units</td>
<td>4. Assigns probabilities to language units</td>
</tr>
</tbody>
</table>

Rationalist: Nobody can manipulate you.
Empirist: It is very easy to manipulate you.
The Turing Test
Modern MT is all about data
Learns to translate from human translations = data

To best support your needs MT needs the best data

Your data = the best data for MT for you

Without the right data it won’t work!
DATA is Everything
The ELRC initiative
The ELRC context
Heads & organisations behind the ELRC

• The ELRC Consortium:
Heads & organisations behind the ELRC

- 30 ELRC Technological National Anchor Points
- 30 ELRC Public Services National Anchor Points
What does ELRC do?

- **Collect**: Language resources
- **Identify**: Needs of public services
- **Engage**: With the public sector in the identification of LR
- **Help**: With any technical or legal issues
- **Act**: Observatory for language resources across Europe
Why ELRC...?

• To facilitate multilingual communication and exchange of information in key public service scenarios
• How? With the right language resources!
• Examples of public services scenarios to which eTranslation may be applied:
  – Consumer Rights (ODR),
  – Social Security (EESSI),
  – Justice (eJustice),
  – Commerce (BRIS),
  – Open Data (EDP),
  – ...

How to make it work...?

In-domain data → Out-of-domain data

→ language resources of the public sector make a difference!
How to make it work...?

- Add in-domain text

CEF AT

- In-domain data

- Good translation

Source Text

Translation
Language Resources of interest

- Translated texts
- Translation memories
- Bi- or multilingual corpora
- Terminologies
What has been achieved so far...?

- Collection of 225 language resources overall
- More than 2 billion words in all EU official languages, Norwegian and Icelandic
- Over 450,000 terms
- More than 2 million translation units
- More than 91 language resources to be used by you!
• Public Sector Information (PSI) Directive 2013/37/EU
  – Amending Directive 2003/98/EC
  – Emphasis on open data
  – Clear cost rules
  – Including Cultural Institutions (Museums, Libraries and Archives) within the scope of the Directive
  – Introduce a clear regime for exclusive agreements
  – Emphasis on standard and machine readable licences
• The ELRC-SHARE Repository
  – Access to, sharing and contribution of language resources
  – Access to tools and services catalogue (upcoming)
  – Visit https://elrc-share.eu
What happens to your data?

Data contributor

Upload to ELRC-SHARE

ELRC processes your data

Processed data
Data processing before delivery to EC

• All datasets are processed to result in tmx/tbx/txt files
• Data will indicatively undergo the following processing:
  – cleaning
  – format conversion
  – sentence alignment
  – metadata completion
• **ELRC Technical and Legal Helpdesk**
  – Continuous support for potential data donors (phone, email, …)
  – Accessible online via [http://www.lr-coordination.eu/helpdesk](http://www.lr-coordination.eu/helpdesk)
• **ELRC On-site assistance**
  
  – Submit request online at [http://www.lr-coordination.eu/services](http://www.lr-coordination.eu/services)
Thank you for your attention!

Email: info@lr-coordination.eu
Website: www.lr-coordination.eu