Machine Translation at the European Commission

Konferenz “10 Jahre Verbmobil”
Saarbrücken, 16. November 2010
This talk is not about MT research, but on (M)T of written documents within a large-scale production environment; yet we try to make use of recent research results as much as possible.

→ Translation at European Commission and other European Institutions
→ Machine Translation at EC
→ MT@EC in Context: Past and Future
Translation at European Commission

Directorate-General for Translation (DGT)

- Staff:
  1750 “linguists” (=translators) and 600 support
- Production (M pages):
- Cost:
  EC translation: 300 M€
  all translation and interpretation: 2€/y per citizen

BUT

to make europa.eu fully multilingual

- translate almost 6,8 million documents
- 8,500 translators working full-time for one year
- infeasible if not using new technologies like MT
How many languages do we translate from and into?

1. Italian
2. German
3. French
4. Dutch
5. English
6. Danish
7. Greek
8. Spanish
9. Portuguese
10. Swedish
11. Finnish
12. Slovenian
13. Slovak
14. Polish
15. Maltese
16. Lithuanian
17. Latvian
18. Hungarian
19. Estonian
20. Czech
21. Romanian
22. Irish *
23. Bulgarian

What do we translate?

- Legal acts and preparatory documents
- Commission decisions and communications
- Publications
- Correspondence
- Speeches, minutes
- Reports, working documents
- Web pages
What do we translate?

Document types at the European Commission

- Existing legal obligations: 74%
- Communication priorities: 8%
- Non-core documents: 10%
- Political priorities / new legal obligations: 8%
Why?

- Council Regulation No 1/58
  - Regulations and other documents of general application shall be drafted in the official languages.

- Treaty establishing the European Community and the Lisbon Treaty
  - Citizens have a right to address the official EU bodies in any of the EU’s official languages and to receive a reply in that language.
Who and how many are we? (# of Translators)

- European Parliament: 760
- European Commission: 1750
- Court of Auditors: 110
- Court of Justice: 620
- Council of the EU: 650
- Committee of the Regions and European Economic and Social Committee: 350
Translation Departments including Support Staff

- 2500 Europäische Kommission Luxemburg – Brüssel
- 1200 Europäisches Parlament, Luxemburg
- 600 Sekretariat des Rates Brüssel
- 600 CITI Interinstitutioneller Übersetzungs- und Dolmetschausschuss
- 530 Ausschuss der Regionen und Wirtschafts- und Sozialausschuss Brüssel
- 200 Übersetzungs­zentrum Luxemburg
- 141 Rechnungshof Luxemburg
- 47 Europäische Investitionsbank Luxemburg
- 851 Gerichtshof Luxemburg
- 65 Europäische Zentralbank Frankfurt

Dolmetschdienste
Rechts- und Sprachsachverständige

Interinstitutionelle Zusammenarbeit im Bereich Übersetzen und Dolmetschen
IT Tools for Partial Automation (CAT)

- ECMT (European Commission Machine Translation)
- Computergestützte Übersetzung (CAT)
- Suche nach Vorgängerübersetzungen
- Multidoc
- EURAMIS
- EUR-Lex

Das Übersetzer-Schaltpult

- Eingabe- und Übersetzungshilfsmittel
- Terminologiehilfsmittel
- Dokumentationshilfsmittel
- Spracherkennung
- Textverarbeitung
- Mehrsprachiges Schreiben
- Optische Zeichenerkennung

- Dicautom
- Euroterm
- Eurodicautom
- TIS
- Euterpe
- Multiterm
- Quest (Metasuchmaschine)
The present MT approach: ECMT service

- managed by DGT
- rule-based machine translation
- developed between 1975 and 1998
- 28 language pairs available (ten languages)
- since 2006 only maintenance reduced work on dictionary enrichment on a couple of systems
- who uses it and what for
  - EU institutions and public bodies for gisting
  - Online services and information systems for raw translation
  - DGT as a CAT tool
The future: MT@EC service

Policy
Commission Communication on "Multilingualism" 2008: “human and automatic translation is an important part of multilingualism policy”

Facts
- ECMT is rule-based and costly to develop
- Data-driven systems are cheap and quick to develop… if you have the data

Language Technology Watch
- Market and research observation
- Tests of commercial and non commercial tools and MT systems
**MT@EC Needs – resources - action**

**MT@EC strategy**
- Adopted in June 2009 by DGT
- Task Force created *November 2009*

**Inter-service Task Force results** *April 2010*
- MT@EC is necessary for the Commission (*trust, confidentiality, continuity*)
- Data-driven systems: a major technological breakthrough
- User requirements have been collected
- An outline of an “architecture” has been elaborated (*flexible, sustainable, ensuring technological independence*)
- Recommendations on organisational and financial arrangements
Machine Translation Service
Outline of the proposed MT@EC architecture

Users and Services

DISPATCHER
- managing MT requests

MT engines
- by language, subject...

MT data
- language resources specific for each MT engine

Language resources
- built around Euramis

Customised interfaces

ENGINES HUB

DATA MODELLING

DATA HUB

USER FEEDBACK
Machine Translation Service
A number of projects within a “MT@EC programme”

“MT Engines - baseline" project (EC)
IT infrastructure for the core of the “MT Engines Hub”

“MT data management hub" projects (DGT)
Language resources (LR) underlying the MT system

“Customised MT solutions" projects (clients)
“Client” requesting development of (examples) :
- a domain specific MT engine
- a specific interface to external services
Exodus

• Internal DGT experimentation with Moses toolkit
• Using Euramis (internal) TM data
• With temporary redeployment of existing ICT and human resources by the DGT ICT unit
• With the active contribution of:
  – the Portuguese language department of the EC
  – the EuromatrixPlus project
  – the Translation DG of the European Parliament
What was done

• Development of the EN->PT engine
• Corpus preparation and cleaning
• Human evaluation by the PT LD (more than 30 translators involved)

What has not been done

(due to time and resource limitations)
• No iterative process for improving corpus quality.
• No incremental updates of translation and language models
• No engineering interventions
First conclusions

• Quality evaluation of MT output for EN->PT results very encouraging

• Dedicated analysis on IT engineering work required for production ready system for all EU languages

• Quality of data cleaning and preparation: the main "comparative" advantage of DGT

Note:
More Exodus pairs are currently being evaluated by the European Parliament, who also submitted an Exodus pair (EN-to-FR) to the WMT 2010 competition
Next: putting pieces together

**MT Action plan** June 2010

- **Action line 1:** MT data
  - **started with:** internal translation memories
  - **challenge:** prepared for optimising all kinds of data for MT

- **Action line 2:** MT engines
  - **started with:** open source tools
  - **challenge:** compare alternative systems (both commercial and non-commercial) in terms of quality of output, price (total cost of ownership), feasibility, language coverage

- **Action line 3:** MT service
  - **started with:** prototype of architecture according to TF
  - **challenge:** flexible and sustainable implementation and governance of MT service

In parallel EC is preparing to continuously update the DGT Multilingual Translation Memory of the Acquis Communautaire (DGT-TM)
A timeline of recent MT activities

• 1988-1993: Fred Jelinek’s team develops SMT at IBM
• 1992-2000: Verbmobil builds hybrid multi-engine MT, including the IBM models
• 1999: JHU summer WS, GIZA released as OS code
• 2000-2010: DARPA/NIST MT campaign
• 2005-2010: WMT workshops & shared tasks
• 2006-2009: EuroMatrix develops MT for EU languages
• 2009-2012: EuroMatrix Plus brings MT for EU languages to the user
• 2010: DGT decides to make use of recent research results for large-scale practical usage

=> Current MT activities at EC can be seen as an indirect result (grand-children) of Verbmobil
… via user feedback

- Many recent activities in SMT research (e.g. WPs in EuroMatrix Plus) explore corrections by end-users to improve SMT models
- Google and other MT providers have been doing this since several years and provide toolkits for web-based post-editing
- These efforts struggle with difficulties to obtain high-quality post-editing results
- Integration of MT into DGT’s workflow gives us access to the daily output of many highly skilled, professional translators
- This will help us to improve our models without extra effort beyond the post-editing that is needed anyhow
- In the long run, this should allow us to go beyond the level of quality offered by today’s MT approaches
Thank you