Recent developments in machine translation policy at the European Patent Office

Dr Georg Artelsmair
Director
European Co-operation
European Patent Office

Brussels, 17 November 2010
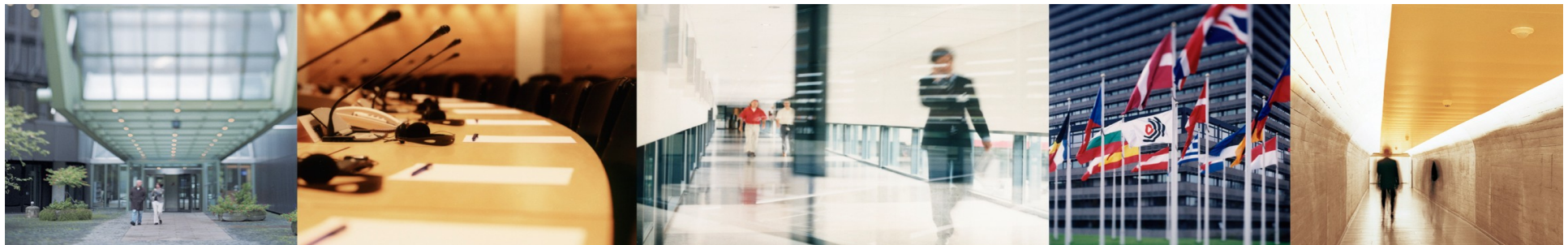
The European Patent Office

Mission

As the patent office for Europe, we support innovation, competitiveness and economic growth across Europe through a commitment to high quality and efficient services delivered under the European Patent Convention.

- Second largest intergovernmental institution in Europe

- Not an EU institution

- Self-financing, i.e. revenue from fees covers operating and capital expenditure

Machine Translation services are relevant to the EPO because they...

- Provide access to patent information to enterprises, researchers and technically qualified users in Europe

- Serve as a contribution to resolving the translation/language issue related to the Community patent

- Support the London Agreement

- Enable examiners to search prior art

The dawn of MT for patents at the EPO: 2004

- Approval of the European Machine Translation Programme (EMTP) by the Administrative Council of the EPO

- Objective: Provide an automated translation service of a sufficient quality to make the technical content of a patent document understandable to a technically qualified person

- Study and Call for tender: only rule based engine bids received

- Quality assessment: EPO selected WorldLingo (using Systran)

- Technical approach used: rule-based engine, hierarchical technical dictionaries built with IPC-based patent terminology

An insight in the creation of technical dictionaries

1. Select, scan and OCR patent documents to acquire matching text in source and target language (NPO & EPO).

2. Align source and target texts on sentence or paragraph level (EPO).

3. Automatically extract terms and their translations from aligned text (external provider).

4. Select term candidates for inclusion in technical dictionaries (EPO).

5. Validate final set of dictionary terms (translation, grammatical information) (external provider).

6. Build bi-directional dictionaries (EPO).

7. Test in Test environment (NPO & EPO).

8. Deploy in Production environment (translation engine provider).

In the meantime...

- 2008: first language pairs, EN-ES/ES-EN and EN-DE/DE-EN, entered into production.

- 2008/9: two further language pairs, EN-FR/FR-EN and EN-IT/IT-EN, entered into production - but improvement still ongoing (quality not satisfactory)

- As per 1 July 2008 IT/EN translation service used for "WOIT" files - enables EPO examiners to carry out prior-art searches and prepare written opinions for Italian files

- 2009: high-quality dictionaries created for SE and PT - interaction with engine delivers poor quality

- 2010: a SMT (Language Weaver) selected for the translation of Italian files due to the persistency of insufficient quality

Some figures

	German		Spanish		French		Italian		Portuguese		Swedish	
	DE-EN	EN-DE	ES-EN	EN-ES	FR-EN	EN-FR	IT-EN	EN-IT	PT-EN	EN-PT	SE-EN	EN-SE
No.documents (5-50 pgs/doc)	871.000		168.046		871.000		108.500		84.885		200.493	
No. created XML files	250.137		42.366		147.972		63.781		32.789		N/A	
No. aligned sentences	7.000.000		5.768.314		4.567.825		6.069.820		3.782.037		N/A	
No. Dictionary terms/words	386.204	332.681	274.979	274.995	213.602	182.933	795.854	764.664	118.071	126.675	1.385.439	1.378.751
Human acceptance score for translation in production	scale (3-9) score: 6		scale (3-9) score: 6		(1-5) score: 4,3	(1-5) score: 3,25	(1-5) score: 2,89	(1-5) score: 2,82	N/A	N/A	N/A	N/A

- The scores for French and Italian language results from EPO internal acceptability test
- In dictionaries the same terms appeared in (for example in 5) different IPC-dictionaries are counted 5 times.
- The score 6 on the scale (3-9) is close to the score 3 on the scale (1-5)

EPO current MT services are available...

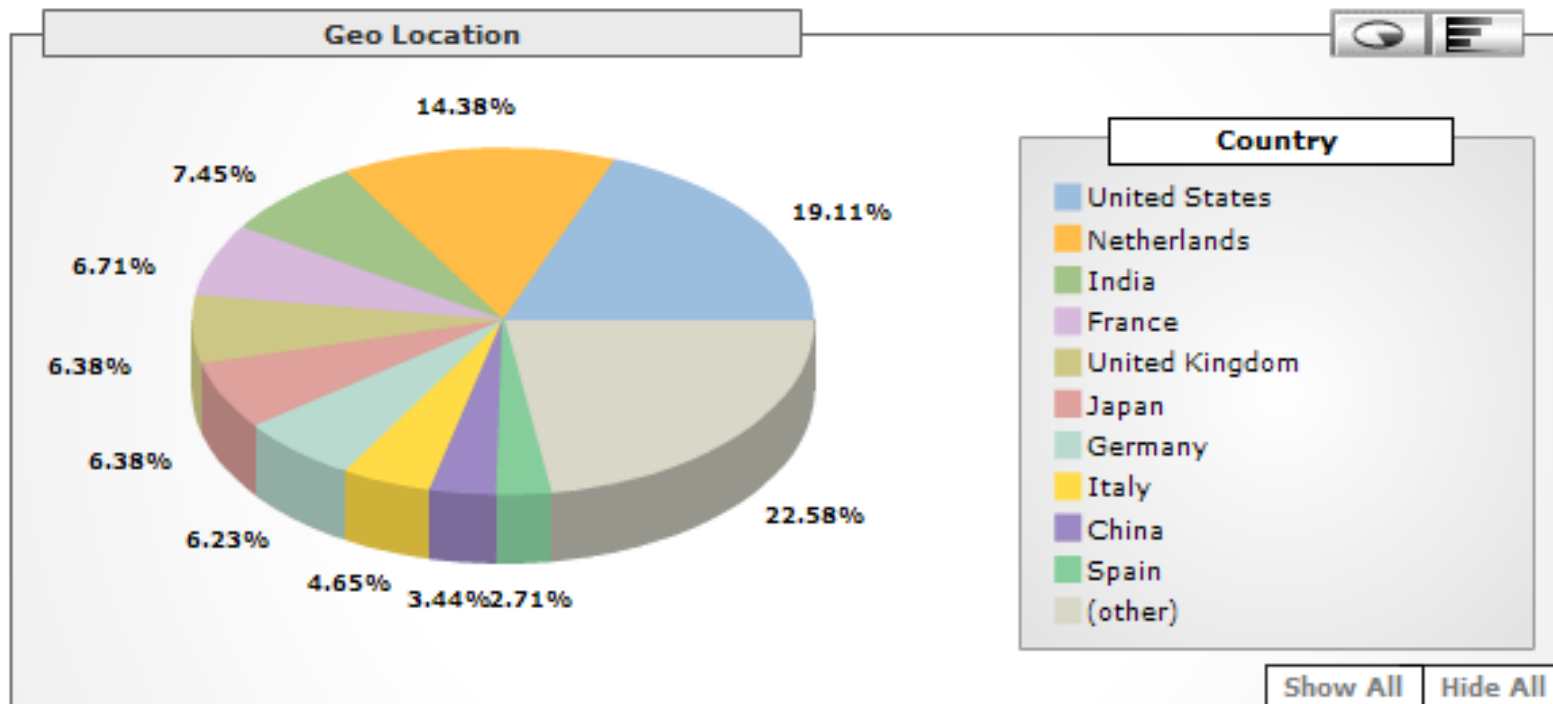- to the public via esp@cenet (abstract, descriptions and claims)

  http://ep.espacenet.com
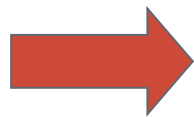
- to the EPO examiners via SEA Viewer from Epoque

Geographical origin of esp@cenet translation requests

# Technical limits of the current approach reached

- Implementation of further language pairs on hold due to:
    - insufficient quality of current engine / technical approach
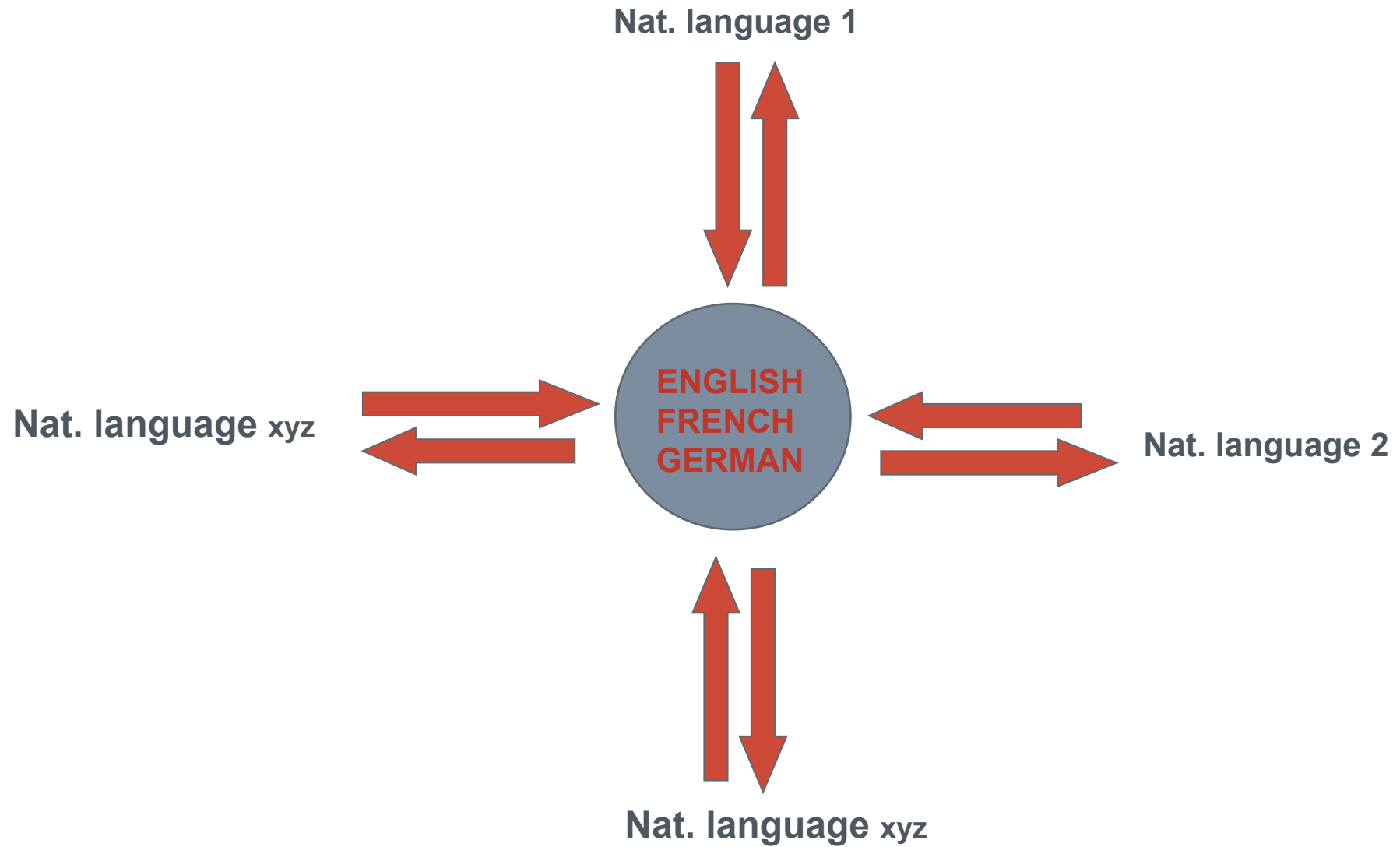   - no suitable rule-based translation engines for certain EPO languages (e.g. RO)

 need to move on to a new concept

What we have today...

Nat. language (DE)

Nat. language xyz

ENGLISH

Nat. language (ES)

Nat. language (IT, SE, FR, PT ...)
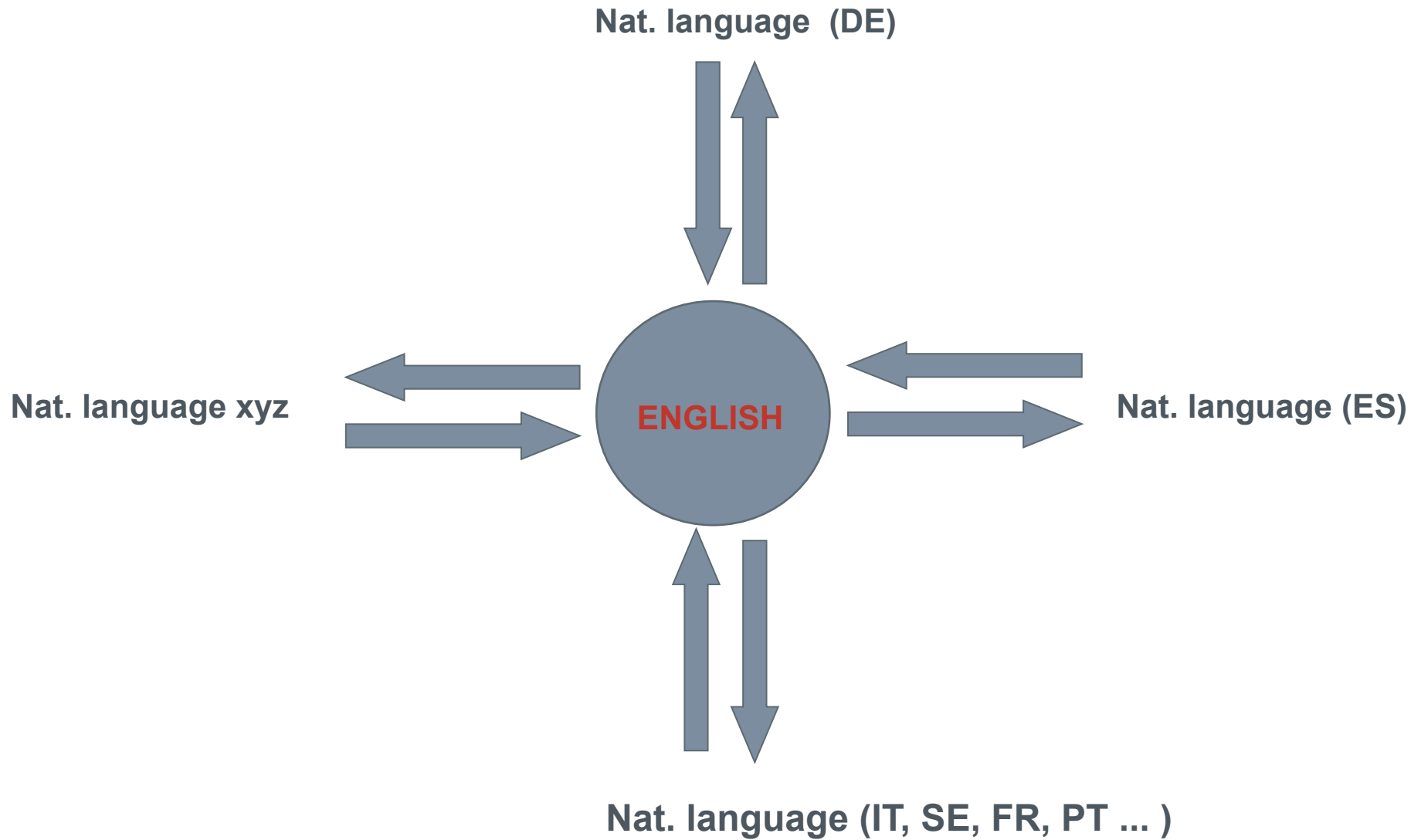
... and what we will need in the future

Europäisches
Patentamt

European
Patent Office

Office européen
des brevets

New programme: European language technology services for patents

- machine translation and, later, other language technology services for patents

- from English (and later on, from French and German)

- into all languages of the EPC contracting states and vice-versa

- to technically qualified users skilled in the art

Objectives

- Support the dissemination of patent information, in particular in the perspective of the forthcoming EU patent

- Support the patent examination procedure

Overall structure

- Phase 1: building corpora of patent documents - collecting of patent documents for enabling the building up of a centralised repository of patent corpora in all EPC contracting states' languages

- Phase 2: language technology services delivery - establishing progressively language technology services for all languages of the EPC contracting states

- Phase 3: integration - intelligent integration of the language technology services into existing tools and services

- Phase 4: maintenance - securing the sustainability and continuous improvement of the services over time

Risks

- Lack of a suitable generic translation engine for each language pair (especially for the translation from and into French and German)

- Lack of patent document pairs (especially for the translation from and into French and German)

The EPO is in contact with the EC in order to identify appropriate solutions and mitigate these risks

Translation quality (Fit-for-purpose)

- Final quality: enable a technically qualified user skilled in the art to understand the technical content of the patent document (fit-for-purpose)

- Service set-up (minimum quality): enable a technically qualified user skilled in the art to assess whether a given patent document is relevant from a technical or economic point of view

Prioritisation

- Automatic translation services from and into English (French and German will follow)

- Languages for which a suitable generic translation engine is available

- Languages for which sufficient patent corpora is available

Role of National Patent Offices

- Provide available national patent documents (at least back to 1990)

- Enable the EPO to treat and use the patent corpora as needed in the programme

- Participate in the quality evaluation

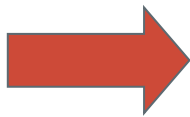- Integrate the services into their websites and tools

Time and budget

- Approval expected at October Admin Council

- Duration: 4 years (Start date: 1 November 2010)

- Budget estimation: 10m € over 4 years

- EPO staff resources: 8-10 m/y (in addition to budget)

And what next?

➢ Growing volume of patent information only available in Asian languages

➡ **automatic translation services for Chinese, Japanese, Korean**

Thank you for your attention