



Europäisches  
Patentamt  
European  
Patent Office  
Office européen  
des brevets

# The Evolution of the European Machine Translation Programme at the European Patent Office

**Patrizia Biani**

Country coordinator

European Affairs/Member States

European Patent Office

Berlin, 7 June 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.

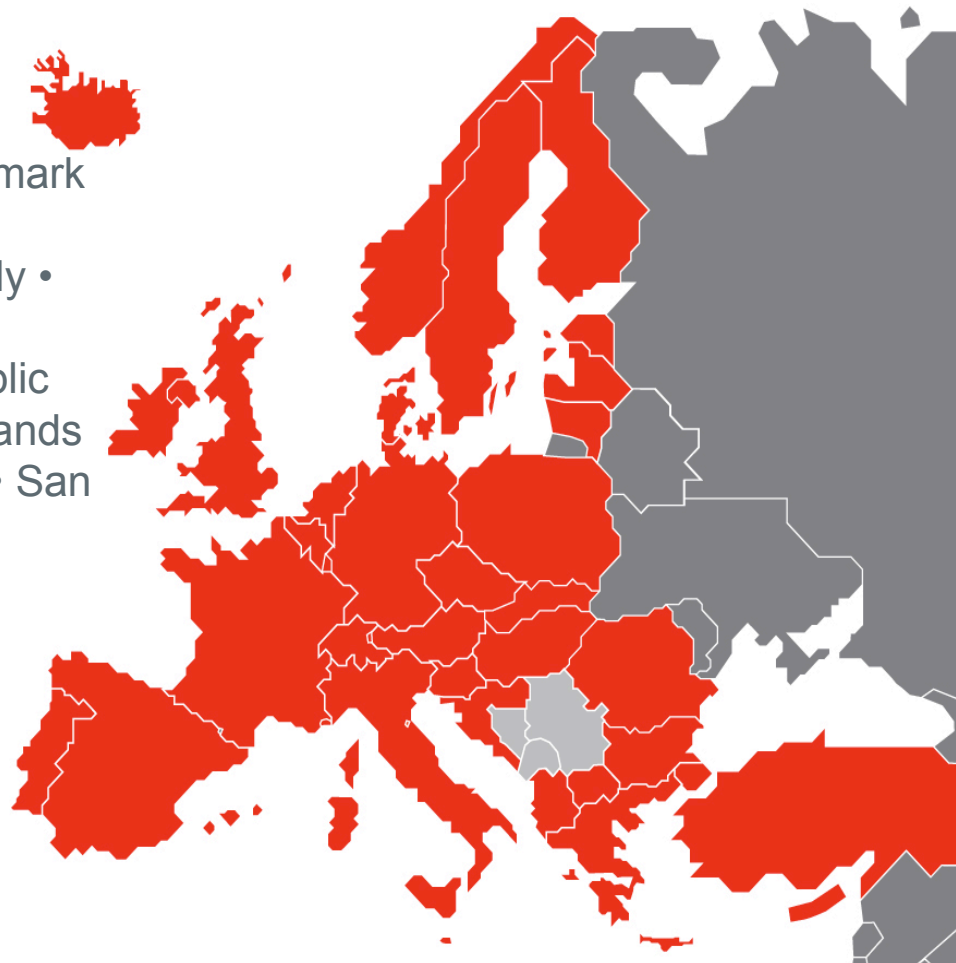
- Second largest intergovernmental institution in Europe
- Not an EU institution
- Self-financing, i.e. revenue from fees covers operating and capital expenditure

## 37 member states

Albania • Austria • Belgium • Bulgaria •  
Croatia • Cyprus • Czech Republic • Denmark  
• Estonia • Finland • France • Germany •  
Greece • Hungary • Iceland • Ireland • Italy •  
Latvia • Liechtenstein • Lithuania •  
Luxembourg • Former Yugoslav Republic  
of Macedonia • Malta • Monaco • Netherlands  
• Norway • Poland • Portugal • Romania • San  
Marino • Slovakia • Slovenia • Spain •  
Sweden • Switzerland • Turkey • United  
Kingdom

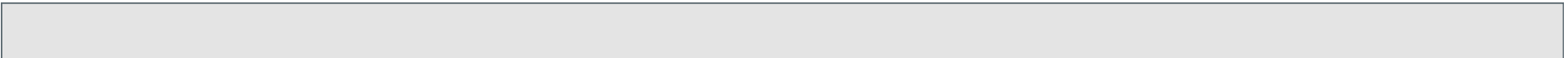
European patent applications and patents  
can also be extended at the applicant's  
request to the following states:

Bosnia-Herzegovina • Montenegro •  
Serbia



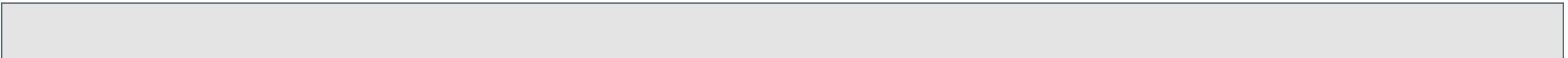
## Relevance of MT services at the EPO

- Provide access to patent information to enterprises, researchers and technically qualified users in Europe
- Support the London Agreement
- Serve as a contribution to resolving the translation/language issue related to the Community patent
- Enable examiners to search prior art



## Historical background

- 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



# 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).

## Some milestones

- 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 - implementations on hold
- 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)

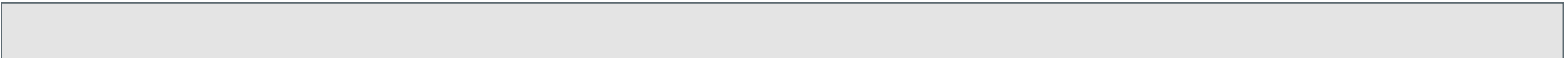


## Availability of EPO MT services

- to the public via   
(translation of abstract, descriptions and claims)

<http://ep.espacenet.com>

- to the EPO examiners via SEA Viewer from Epoque



## LARGE LANGUAGE MODELS IN MACHINE TRANSLATION

### Bibliographic data

Description

Claims

Mosaics

Original document

INPADOC legal status

**Publication number:** KR20100015518 (A)

**Publication date:** 2010-02-12

**Inventor(s):** BRANTS THORSTEN [DE]; POPAT ASHOK C [US]; XU PENG [CN]; OCH FRANZ JOSEF [DE]; DEAN JEFFREY [US] +

**Applicant(s):** GOOGLE INC [US] +

#### Classification:


- international: G06F17/28; G10L15/18; G06F17/28; G10L15/00


- European: G06F17/28D2; G06F17/28D4; G06F17/28D8


**Application number:** KR20097021287 20080325


**Priority number(s):** US20070767436 20070622; US20070920283P 20070326

#### Also published as:

 WO2008118905 (A2)

 WO2008118905 (A3)

 US2008243481 (A1)

 EP2137639 (A2)

[View INPADOC patent family](#)

[View list of citing documents](#)

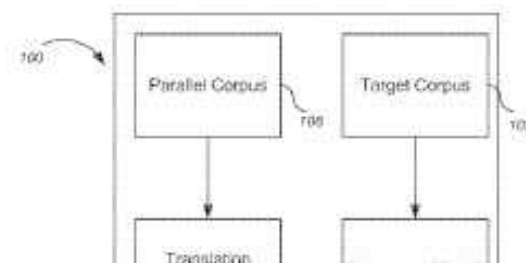
[Report a data error here](#)

Abstract not available for KR 20100015518 (A)

Abstract of corresponding document: **WO 2008118905 (A2)**

[Translate this text](#)

Systems, methods, and computer program products for machine translation are provided. In some implementations a system is provided. The system includes a language model including a collection of n-grams from a corpus, each n-gram having a corresponding relative frequency in the corpus and an order n corresponding to a number of tokens in the n-gram, each n-gram corresponding to a backoff n-gram having an order of n-1 and a collection of backoff scores, each backoff score associated with an n-gram, the backoff score determined as a function of a



## LARGE LANGUAGE MODELS IN MACHINE TRANSLATION

Bibliographic data

Description

Claims

Mosaics

Original document

INPADOC legal status

The EPO does not accept any responsibility for the accuracy of data and information originating from other authorities than the EPO; in particular, the EPO does not guarantee that they are complete, up-to-date or fit for specific purposes.

Description not available for **KR 20100015518 (A)**

Description of corresponding document: **WO 2008118905 (A2)**

Translate this text

LARGE LANGUAGE MODELS IN MACHINE TRANSLATION

### BACKGROUND

This specification relates to statistical machine translation.

Manual translation of text by a human operator can be time consuming and costly. One goal of machine translation is to automatically translate text in a source language to corresponding text in a target language. There are several different approaches to machine translation including example-based machine translation and statistical machine translation. Statistical machine translation attempts to identify a most probable translation in a target language given a particular input in a source language. For example, when translating a sentence from French to English, statistical machine translation identifies the most probable English sentence given the French sentence. This maximum likelihood translation can be written as:  $\arg \max P(e | f)$  which describes the English sentence,  $e$ , out of all possible sentences, that provides the highest value for  $P(e | f)$ . Additionally, Bayes Rule provides that:

$P(e)P(f | e)$

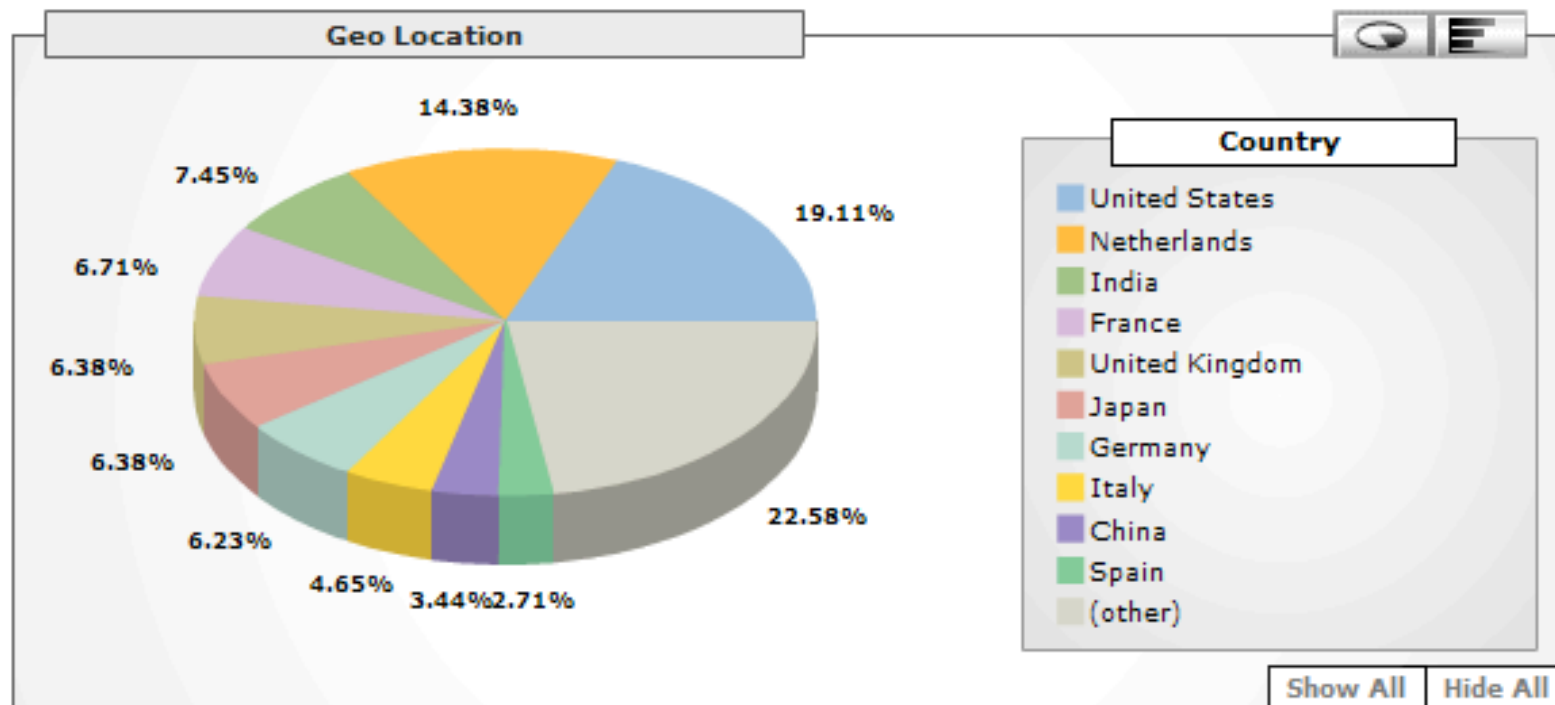


## Number of translation requests (Jan-Apr 2010) ...

- ES → EN: ca 1.800
- EN → ES: ca 33.000
  
- DE → EN: ca 127.000
- EN → DE: ca 20.000
  
- FR → EN: ca 52.000
- EN → FR: ca 50.000
  
- IT → EN: ca 3.500
- EN → IT: ca 20.000

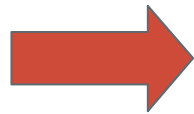
## ... and their geographical origin

### Countries



## Technical limits of the current approach reached

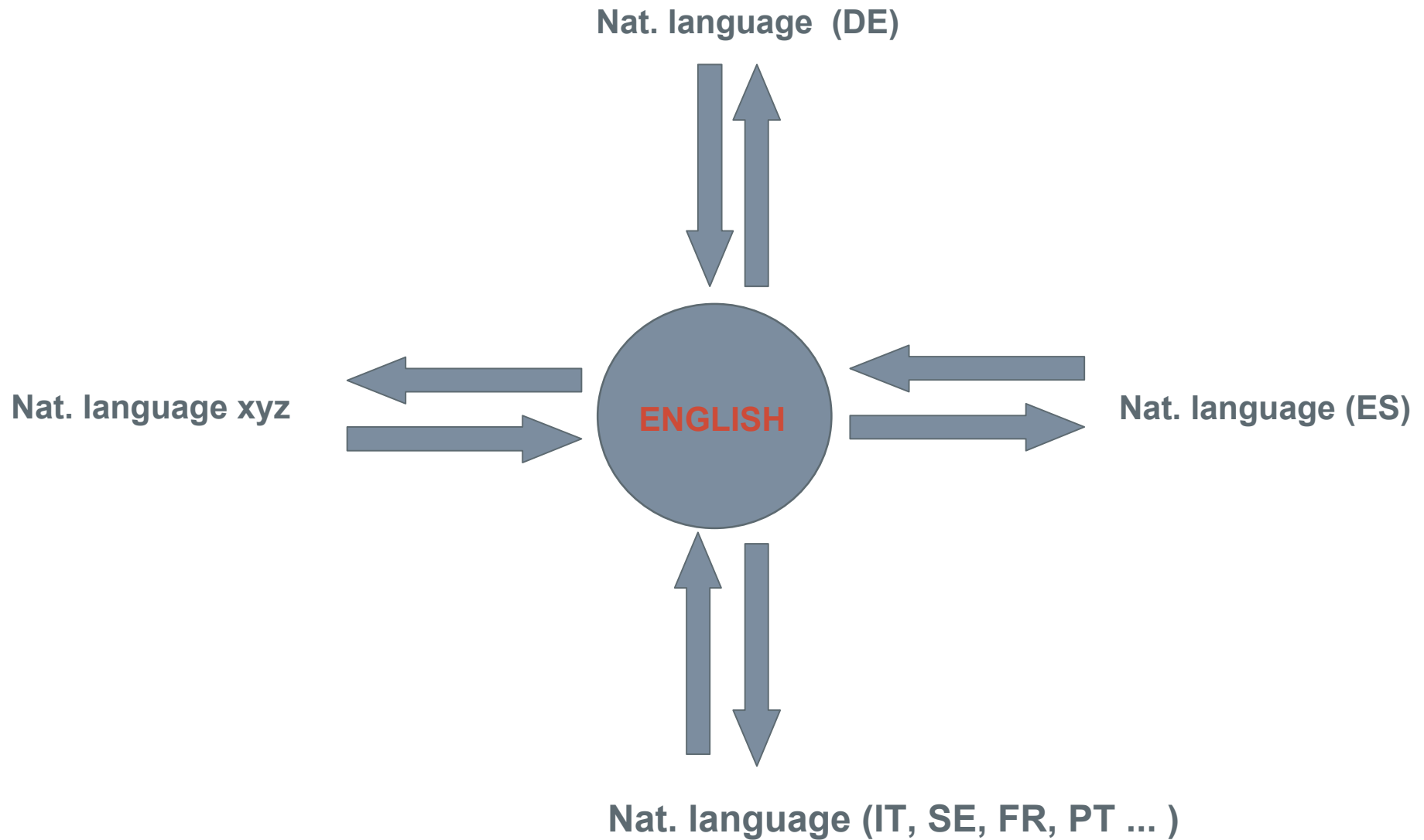
- Implementation of further language pairs on hold due to:
  - insufficient quality of current engine
  - unsatisfactory interaction dictionaries / rule-based engine
  - no suitable rule-based translation engines for certain EPO languages



**need to move on to a new concept**

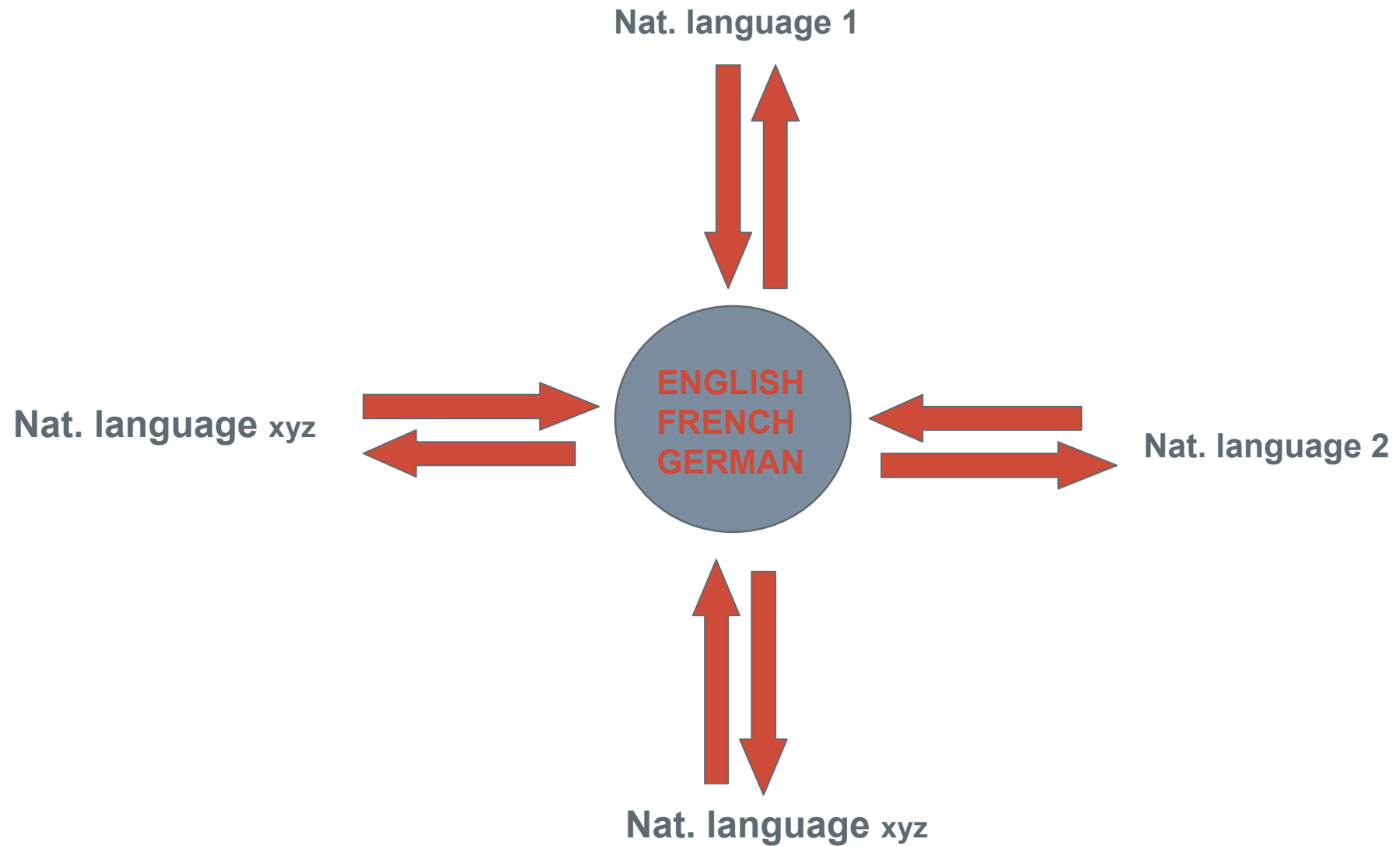
- Identify the most suitable MT technologies and approaches in order to offer, internally and externally, MT solutions which:
  - are tailor-made for translating patent documents providing the necessary accuracy
  - offer scalable machine translation services from the three EPO official languages into **ALL** member states' languages

## What we have today...



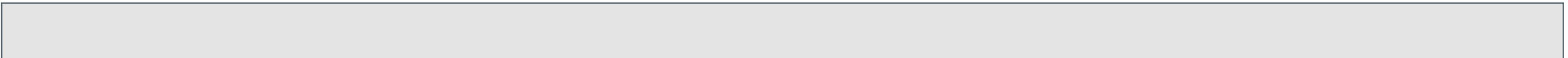


## ... and what we will need in the future



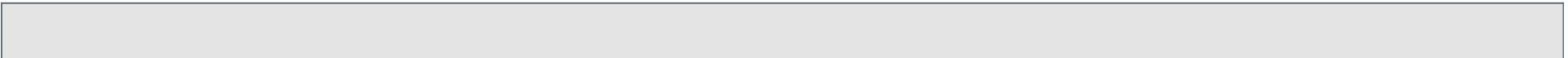
## Result: a new MT co-operation programme

- a new comprehensive MT programme is drafted
- to be approved by the EPO Administrative Council (objective: October 2010)
- within the co-operation framework existing between EPO and member states
  - it serves the users of the European patent system
  - it needs broad support and effort
- in parallel collaboration with EU consortium



## Elements of the new MT co-operation programme

- Contribution from the national patent offices of the member states (patent documents)
- MT technology monitoring (interaction with leading MT research groups and business entities)
- Possibility to use more providers in parallel, also for the same language pair
- Application of effective methods for identification of best suited technology solution for a particular language pair
- Exploring of quality enhancement measures
- Integration of the MT service into other patent tools and services



**Thank you for your attention**

