Invited Talk 1

Machine Translation Project and Activities in EPO
“The patent case”

Mr. Bertrand Le Chapelain
European Patent Office (EPO)

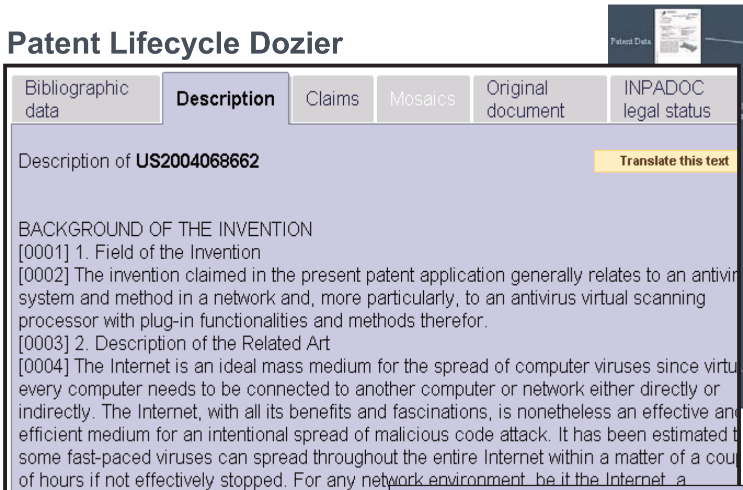
Machine Translation Project and Activities in EPO

"The patent case"

September 2011
EPO, Le Chapelain

Patent Environment

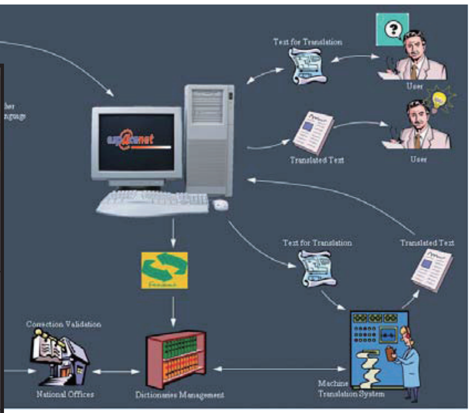
Patent Lifecycle Dozier



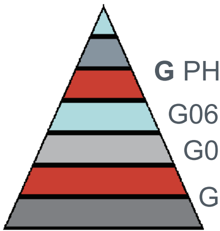
Description of **US2004068662**

BACKGROUND OF THE INVENTION
[0001] 1. Field of the Invention
[0002] The invention claimed in the present patent application generally relates to an antivir
system and method in a network and, more particularly, to an antivirus virtual scanning processor with plug-in functionalities and methods therefor.
[0003] 2. Description of the Related Art
[0004] The Internet is an ideal mass medium for the spread of computer viruses since virtu
every computer needs to be connected to another computer or network either directly or indirectly. The Internet, with all its benefits and fascinations, is nonetheless an effective anc
efficient medium for an intentional spread of malicious code attack. It has been estimated t
some fast-paced viruses can spread throughout the entire Internet within a matter of a cou
of hours if not effectively stopped. For any network environment, be it the Internet, a metropolitan area network (MAN), a wide area
even wireless communications networks for mc
(PDA) devices, the more data transmitted and t
viruses are able to infect those networks.

Patent IT and Process



International Patent Classification (IPC):



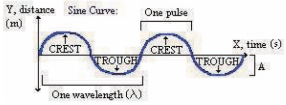
G PHYSICS (Biology, Chemistry, …)
G06 COMPUTING CALCULATING COUNTING
G06F ELECTRIC DIGITAL DATA PROCESSING
G06F19 Digital computing

Representative Problem

Objective: To provide automated translation services to make the technical content of a patent document sufficiently understandable for a technically qualified person.
Aim: to solve ambiguity within one IPC and among different IPCs.

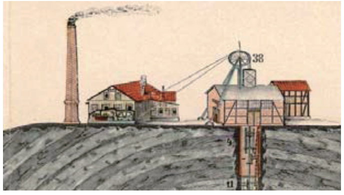
Translation into EN of the DE lexeme ***Welle***:

- ***Wave*** in the domains of transmission system, water engineering.

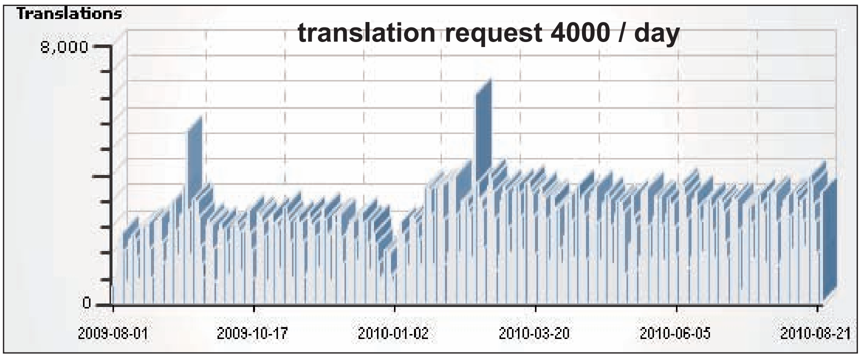


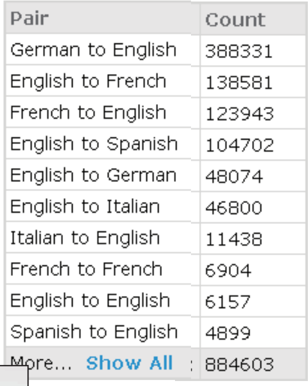
- ***Shaft*** in the domain of mechanical engineering.

- Translation into DE of the EN lexeme *shaft:*
  - "**Schacht**" in the domain of Mining



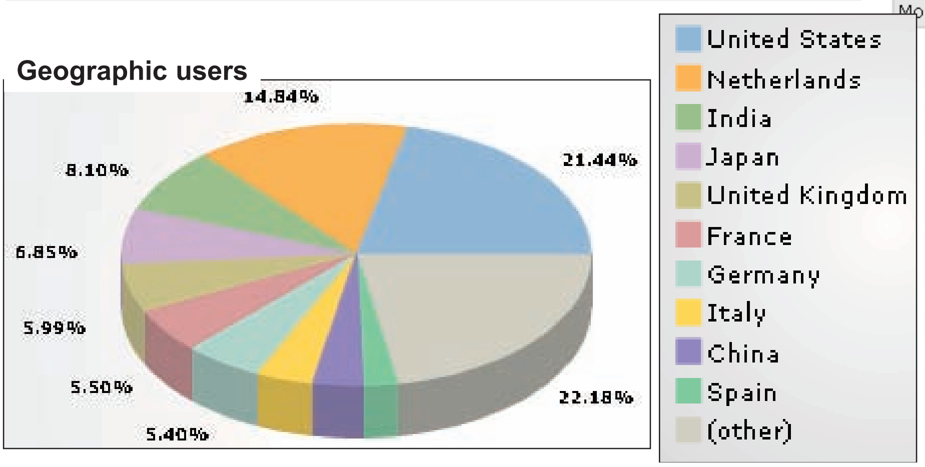
Machine Translation in daily usage





Pair	Count
German to English	388331
English to French	138581
French to English	123943
English to Spanish	104702
English to German	48074
English to Italian	46800
Italian to English	11438
French to French	6904
English to English	6157
Spanish to English	4899
More... Show All :	884603

Geographic users



Main Requirements

Translation for language groups:

- **27 languages** to EN, FR, DE
- **160 Language - pairs**: DE, ES, FR, IT, PT, SV, FI, EL, NL, RO, CS, DA, HU, PL, SK, BG,
  ET, LT, LV, SL, SQ, HR, IS, MK, NO, SR, TR

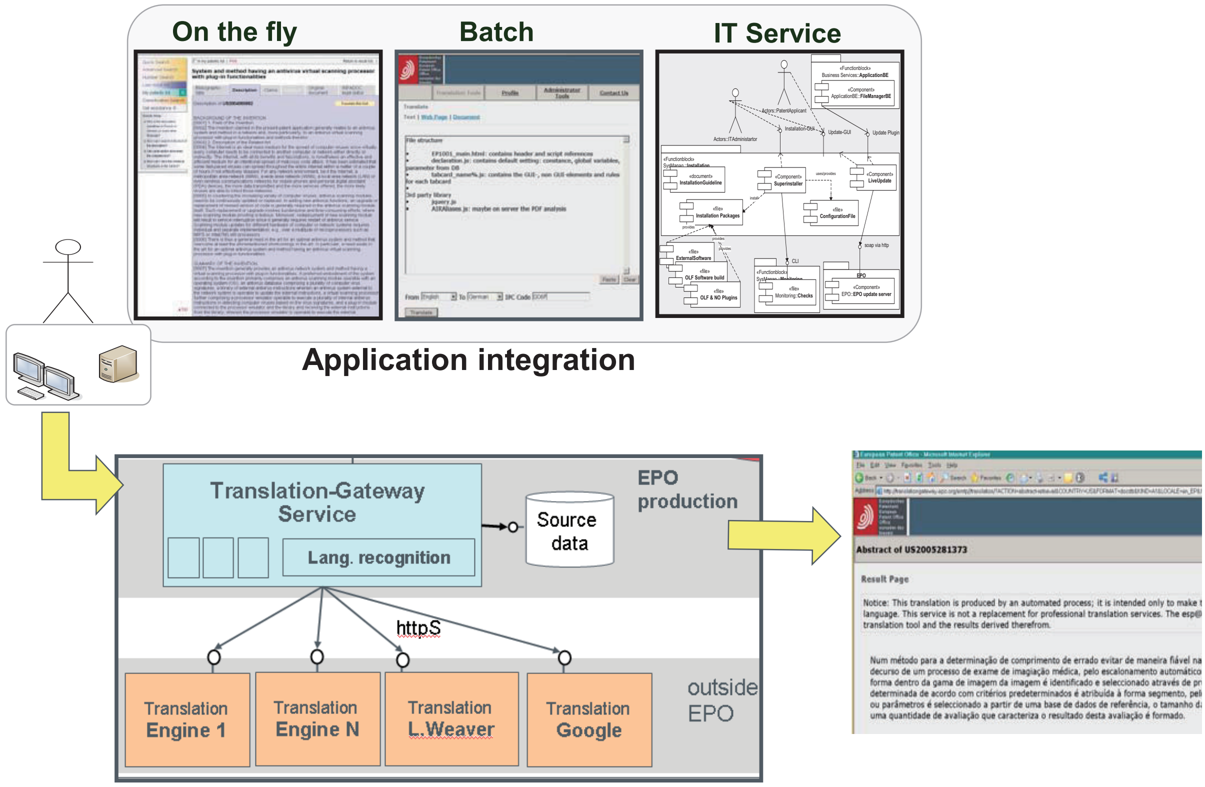
Translation quality

- **Final quality**: enable a technically qualified user skilled in the art to understand the technical content of the patent document
- **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

User / Application:

- Support the dissemination in the perspective of the forthcoming **EU patent**
- The integration into the **patent information** applications
- Support the patent **examination procedure** and **national patent offices** applications

Solution Concept



File structure

- EP1001_main.html: contains header and script references
- declaration.js: contains default setting: constance, global variables, parameter from DB
- tabcard_name%.js: contains the GUI-, non GUI-elements and rules for each tabcard
-

3rd party library
- jquery.js
- AIRAliases.js: maybe on server the PDF analysis

From English To German IPC Code [illegible]

European Patent Office - Microsoft Internet Explorer

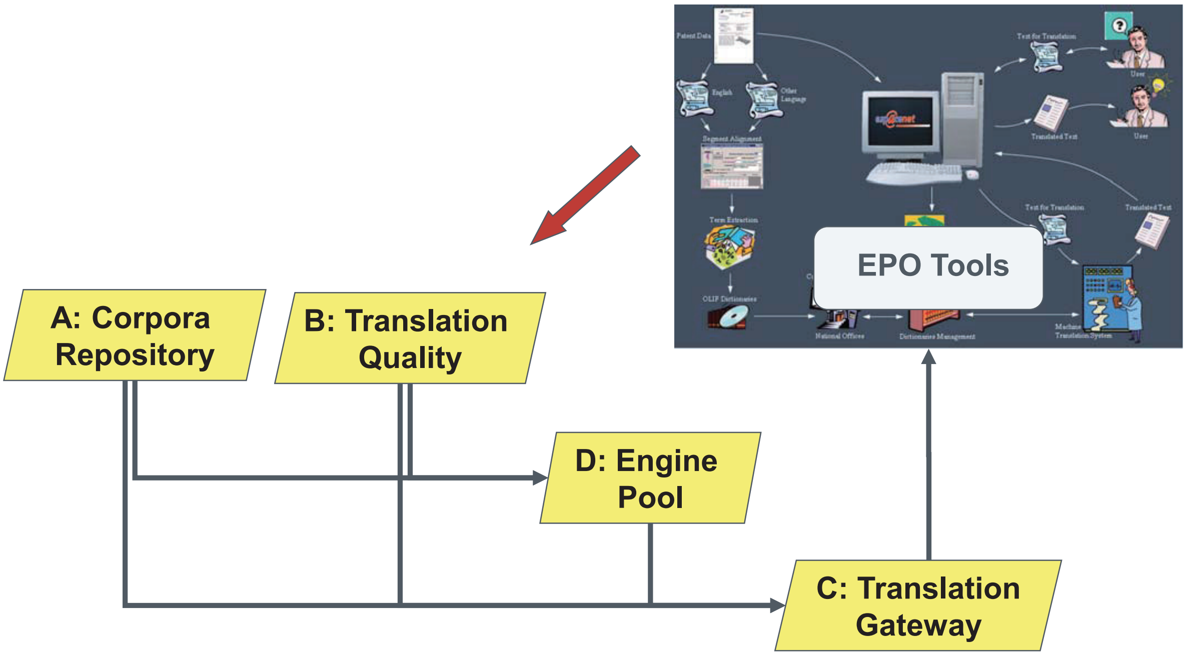
Abstract of US2005281373

Result Page

Notice: This translation is produced by an automated process; it is intended only to make t
language. This service is not a replacement for professional translation services. The esp@
translation tool and the results derived therefrom.

Num método para a determinação de comprimento de errado evitar de maneira fiável na
decurso de um processo de exame de imagiação médica, pelo escalonamento automático
forma dentro da gama de imagem da imagem é identificado e seleccionado através de pr
determinada de acordo com critérios predeterminados é atribuída à forma segmento, pel
ou parâmetros é seleccionado a partir de uma base de dados de referência, o tamanho d
uma quantidade de avaliação que caracteriza o resultado desta avaliação é formado.

Solution Decomposition



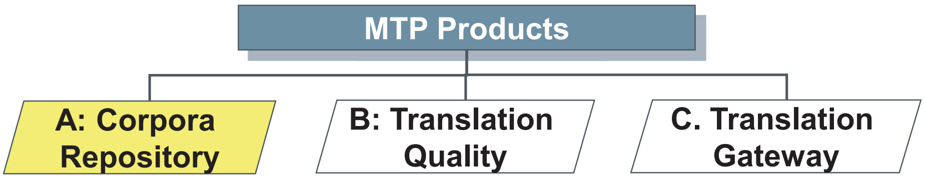
EPO Tools

A: Corpora Repository

B: Translation Quality

D: Engine Pool

C: Translation Gateway



MTP Products

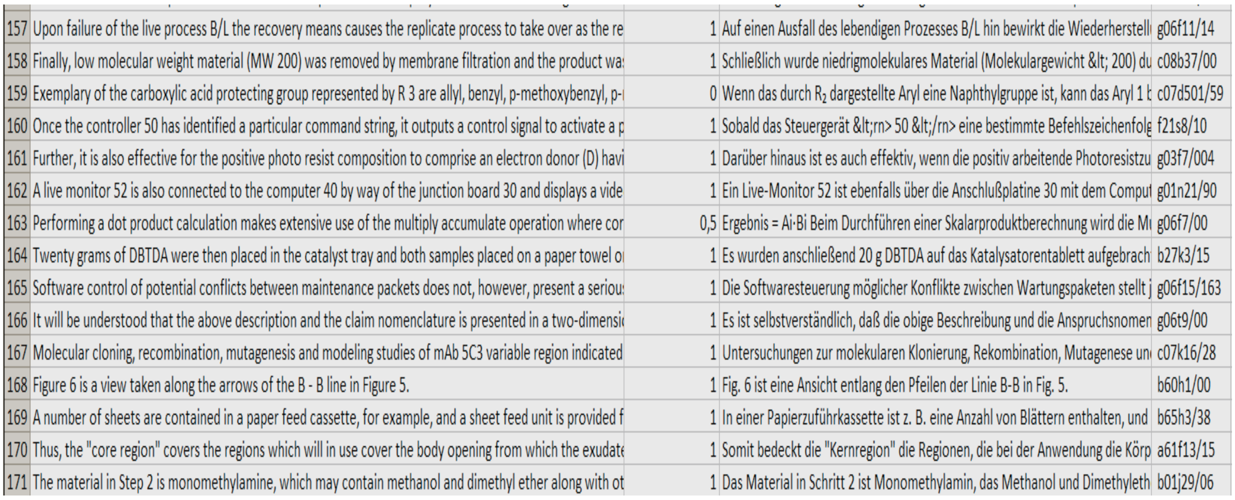
A: Corpora Repository

B: Translation Quality

C. Translation Gateway

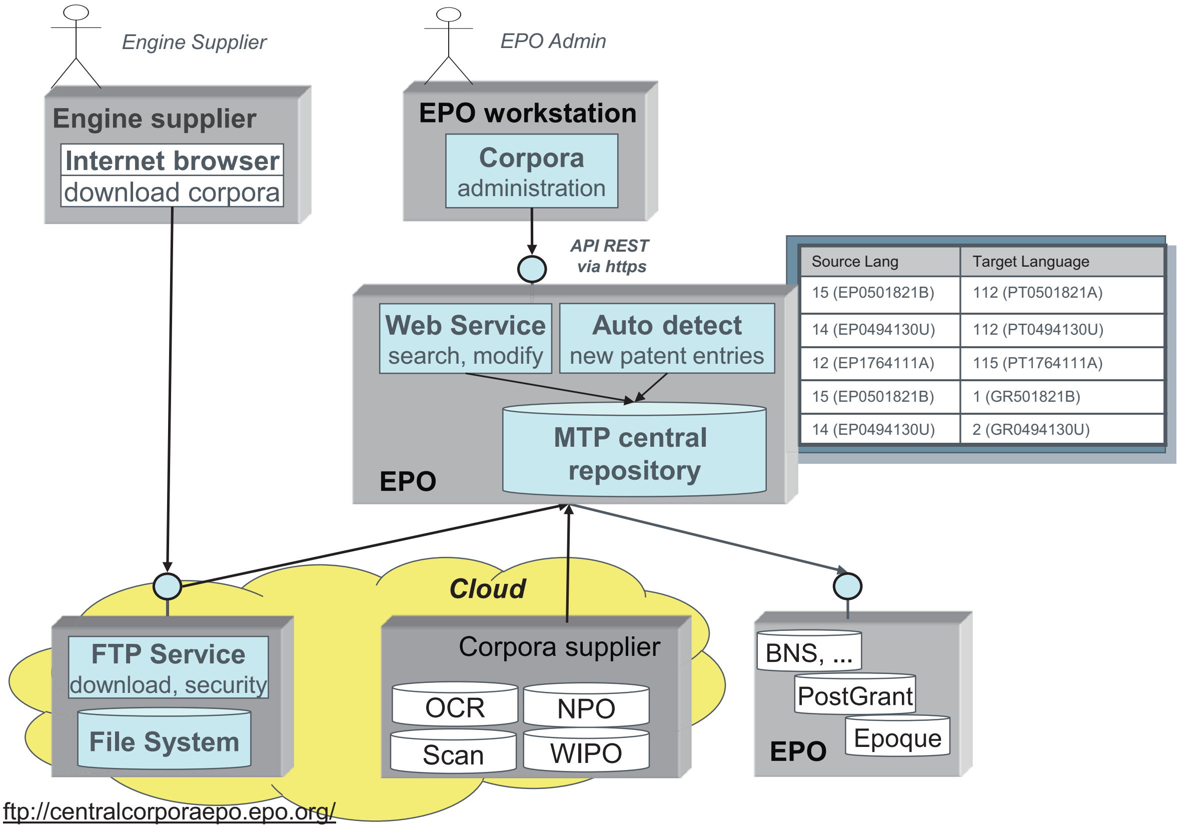
Sentence Alignment: as quality check

Measurement of quality: **Does it match 100%, 75%, 50%, 25%, 0% ???**



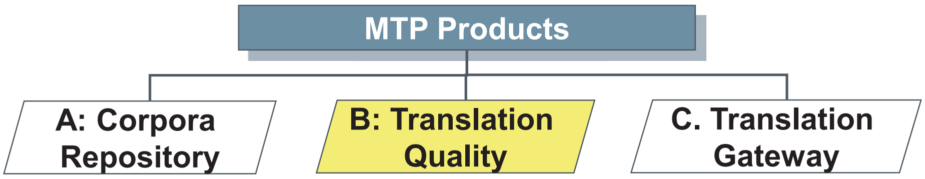
157	Upon failure of the live process B/L the recovery means causes the replicate process to take over as the re	1	Auf einen Ausfall des lebendigen Prozesses B/L hin bewirkt die Wiederherstell	g06f11/14
158	Finally, low molecular weight material (MW 200) was removed by membrane filtration and the product wa:	1	Schließlich wurde niedrigmolekulares Material (Molekulargewicht < 200) du	c08b37/00
159	Exemplary of the carboxylic acid protecting group represented by R 3 are allyl, benzyl, p-methoxybenzyl, p-	0	Wenn das durch R_2 dargestellte Aryl eine Naphthylgruppe ist, kann das Aryl 1 b	c07d501/59
160	Once the controller 50 has identified a particular command string, it outputs a control signal to activate a p	1	Sobald das Steuergerät <rn> 50 </rn> eine bestimmte Befehlszeichenfolg	f21s8/10
161	Further, it is also effective for the positive photo resist composition to comprise an electron donor (D) havi	1	Darüber hinaus ist es auch effektiv, wenn die positiv arbeitende Photoresistzu	g03f7/004
162	A live monitor 52 is also connected to the computer 40 by way of the junction board 30 and displays a vide	1	Ein Live-Monitor 52 ist ebenfalls über die Anschlußplatine 30 mit dem Comput	g01n21/90
163	Performing a dot product calculation makes extensive use of the multiply accumulate operation where cor	0,5	Ergebnis = Ai·Bi Beim Durchführen einer Skalarproduktberechnung wird die Mı	g06f7/00
164	Twenty grams of DBTDA were then placed in the catalyst tray and both samples placed on a paper towel oı	1	Es wurden anschließend 20 g DBTDA auf das Katalysatorentablett aufgebrach	b27k3/15
165	Software control of potential conflicts between maintenance packets does not, however, present a seriou:	1	Die Softwaresteuerung möglicher Konflikte zwischen Wartungspaketen stellt j	g06f15/163
166	It will be understood that the above description and the claim nomenclature is presented in a two-dimensi	1	Es ist selbstverständlich, daß die obige Beschreibung und die Anspruchsnomen	g06t9/00
167	Molecular cloning, recombination, mutagenesis and modeling studies of mAb 5C3 variable region indicated	1	Untersuchungen zur molekularen Klonierung, Rekombination, Mutagenese un	c07k16/28
168	Figure 6 is a view taken along the arrows of the B - B line in Figure 5.	1	Fig. 6 ist eine Ansicht entlang den Pfeilen der Linie B-B in Fig. 5.	b60h1/00
169	A number of sheets are contained in a paper feed cassette, for example, and a sheet feed unit is provided f	1	In einer Papierzuführkassette ist z. B. eine Anzahl von Blättern enthalten, und	b65h3/38
170	Thus, the "core region" covers the regions which will in use cover the body opening from which the exudate	1	Somit bedeckt die "Kernregion" die Regionen, die bei der Anwendung die Körp	a61f13/15
171	The material in Step 2 is monomethylamine, which may contain methanol and dimethyl ether along with ot	1	Das Material in Schritt 2 ist Monomethylamin, das Methanol und Dimethyleth	b01j29/06

Corpora Repository



Source Lang	Target Language
15 (EP0501821B)	112 (PT0501821A)
14 (EP0494130U)	112 (PT0494130U)
12 (EP1764111A)	115 (PT1764111A)
15 (EP0501821B)	1 (GR501821B)
14 (EP0494130U)	2 (GR0494130U)

ftp://centralcorporaepo.epo.org/

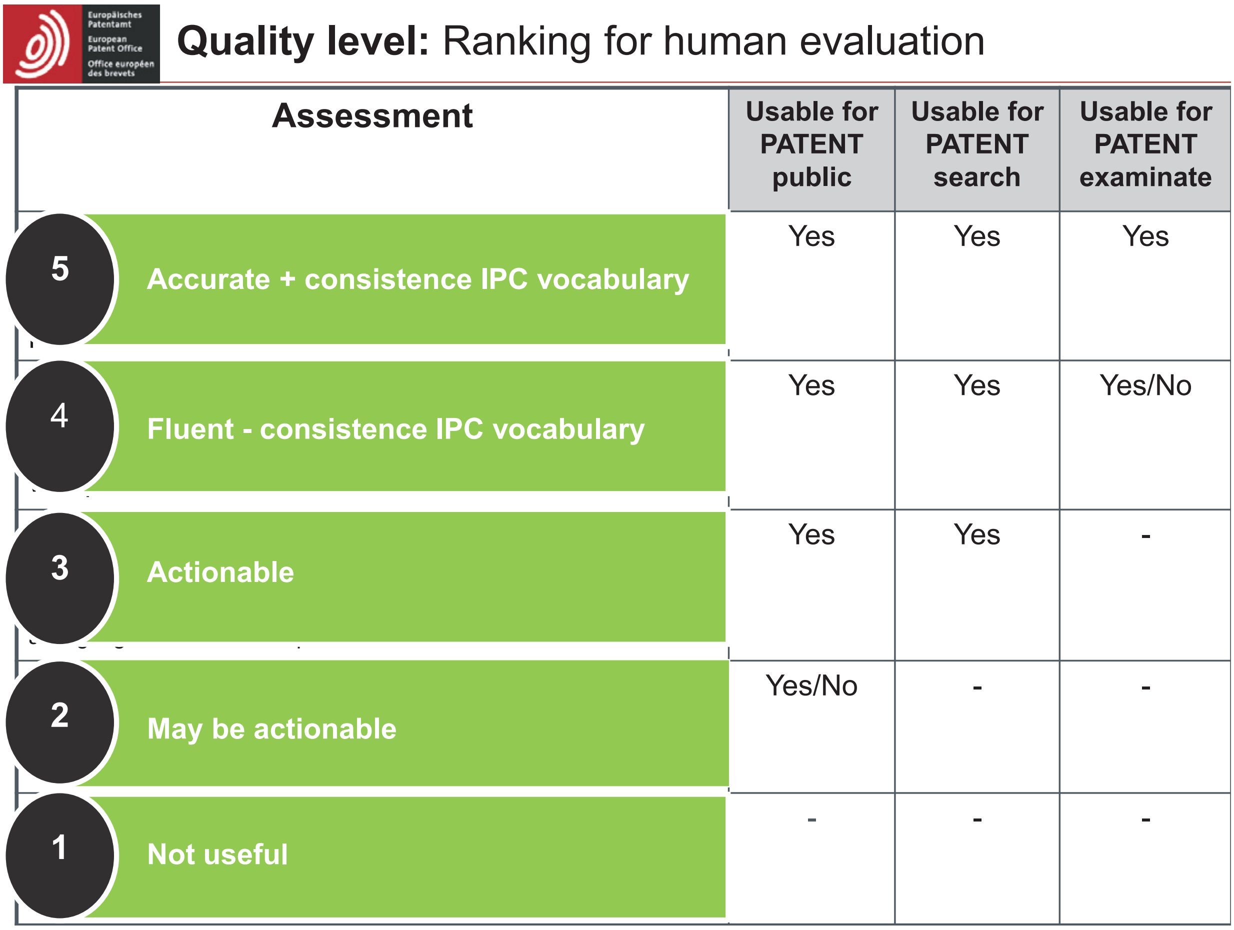


MTP Products

A: Corpora Repository

B: Translation Quality

C. Translation Gateway

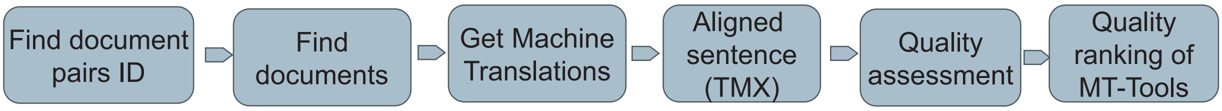


Quality level: Ranking for human evaluation

Assessment	Usable for PATENT public	Usable for PATENT search	Usable for PATENT examinate
5 Accurate + consistence IPC vocabulary	Yes	Yes	Yes
4 Fluent - consistence IPC vocabulary	Yes	Yes	Yes/No
3 Actionable	Yes	Yes	-
2 May be actionable	Yes/No	-	-
1 Not useful	-	-	-

Workflow of the automatic MT Quality evaluation

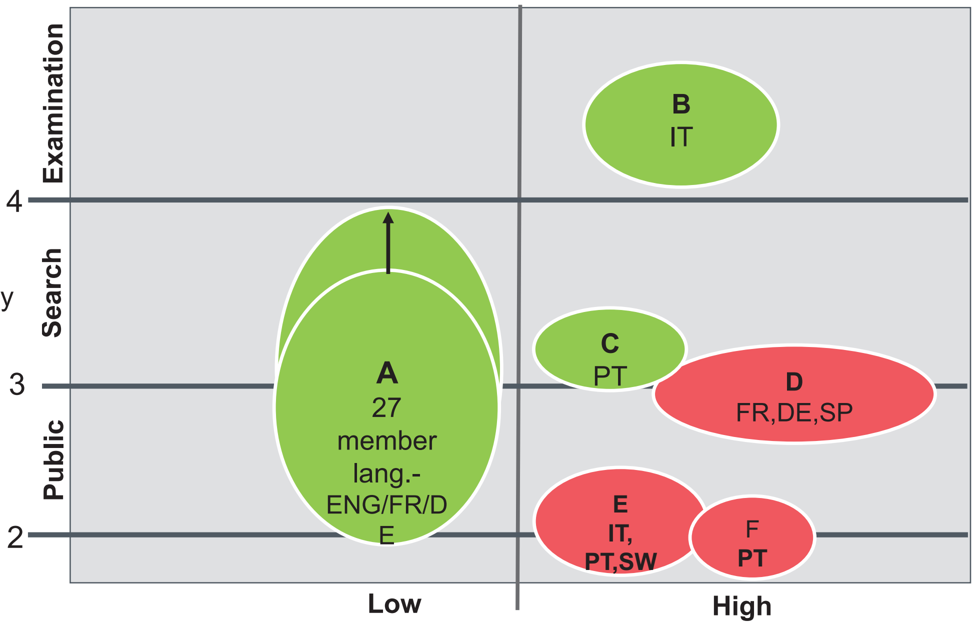
1. Identify document pairs (available in oriental language and English as original or human translation)
2. Find documents
3. Generate or get machine translations from all sources to be tested
4. Align sentences
5. Perform quality assessment
6. Establish quality ranking



Engine Score / Portfolio

Impression (Humans)

- Fluency
- Comprehensibility
- Time to read
- Readability
- Grammar
- Accuracy



Accuracy (Automated quality measurement)

- translated words coverage
- terms correctness IPC
- special characters, formulae, chemical elements, format

Questionnaire for Human Score

Readability: how easy it is to read the text in general

Translation of the whole text
Fluency
Sentence Structure
Text separation

Comprehensibility: the extent to which the meaning of the text is understandable

Translation of technical parts of the text
Translation of the most common language in the text

Overall impression: was it easy to read and to understand

The translation was in general a good translation
It was not a problem to read the text
I have a good unders
It took me more or les
as any other text I no

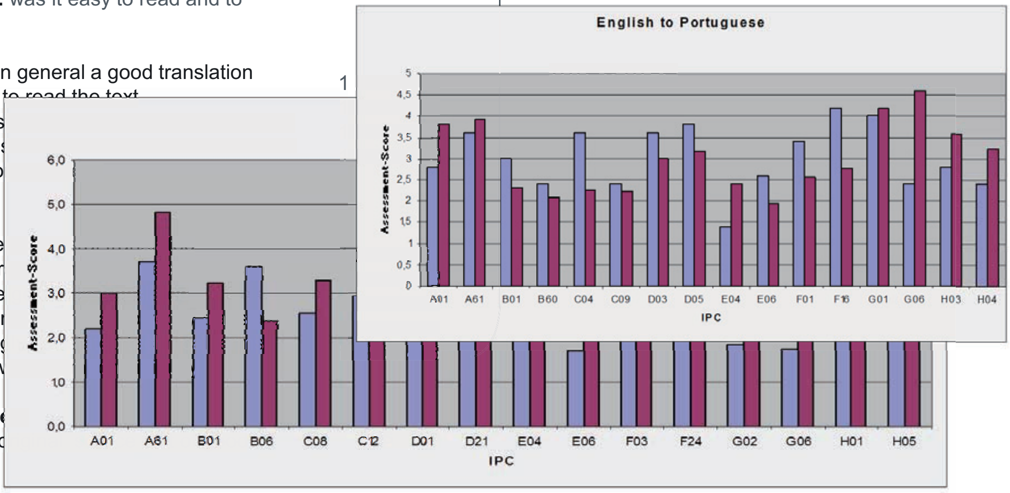
Word Translation
Technical terms were
Translated terms wer
Common words were
Translated words wer
Despite the not/wrong
the text can be read v

Quality of original te
Was the style of the o
and readable quality

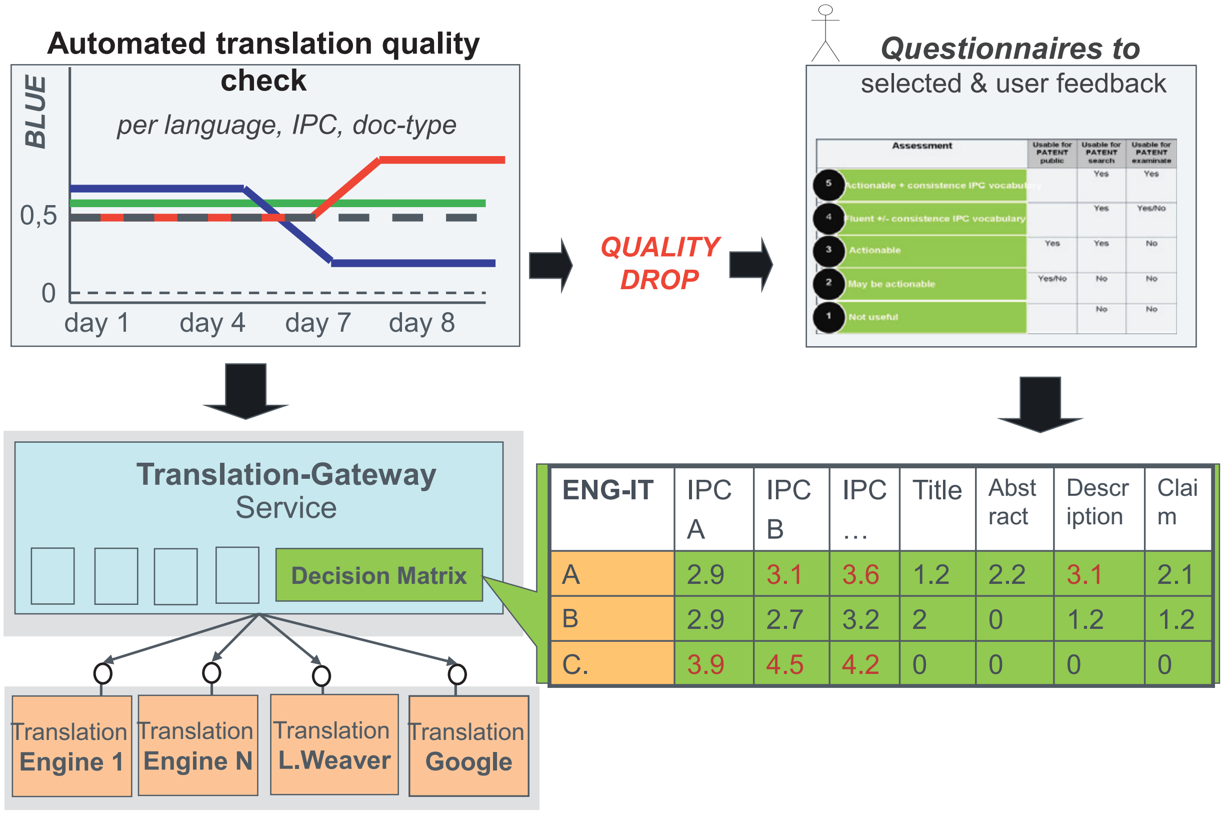
1 2 3 4 5
1 2 3 4 5

ENG-IT	IPC A	IPC B	IP C	Title	Abstract	Description	Claim
A	2.9	3.1	3.6	1.2	2.2	3.1	2.1
B	2.9	2.7	3.2	2	0	1.2	1.2
C	3.9	4.5	4.2	0	0	0	0

1

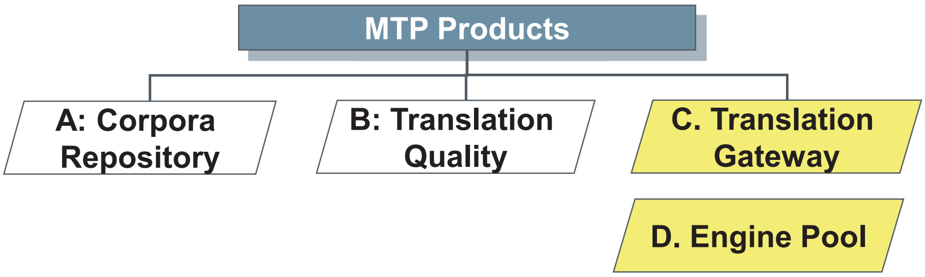


Quality Process in daily usage



Assessment	Usable for PATENT public	Usable for PATENT search	Usable for PATENT examinate
5 Actionable + consistence IPC vocabulary		Yes	Yes
4 Fluent +/- consistence IPC vocabulary		Yes	Yes/No
3 Actionable	Yes	Yes	No
2 May be actionable	Yes/No	No	No
1 Not useful		No	No

ENG-IT	IPC A	IPC B	IPC …	Title	Abstract	Description	Claim
A	2.9	3.1	3.6	1.2	2.2	3.1	2.1
B	2.9	2.7	3.2	2	0	1.2	1.2
C.	3.9	4.5	4.2	0	0	0	0



MTP Products

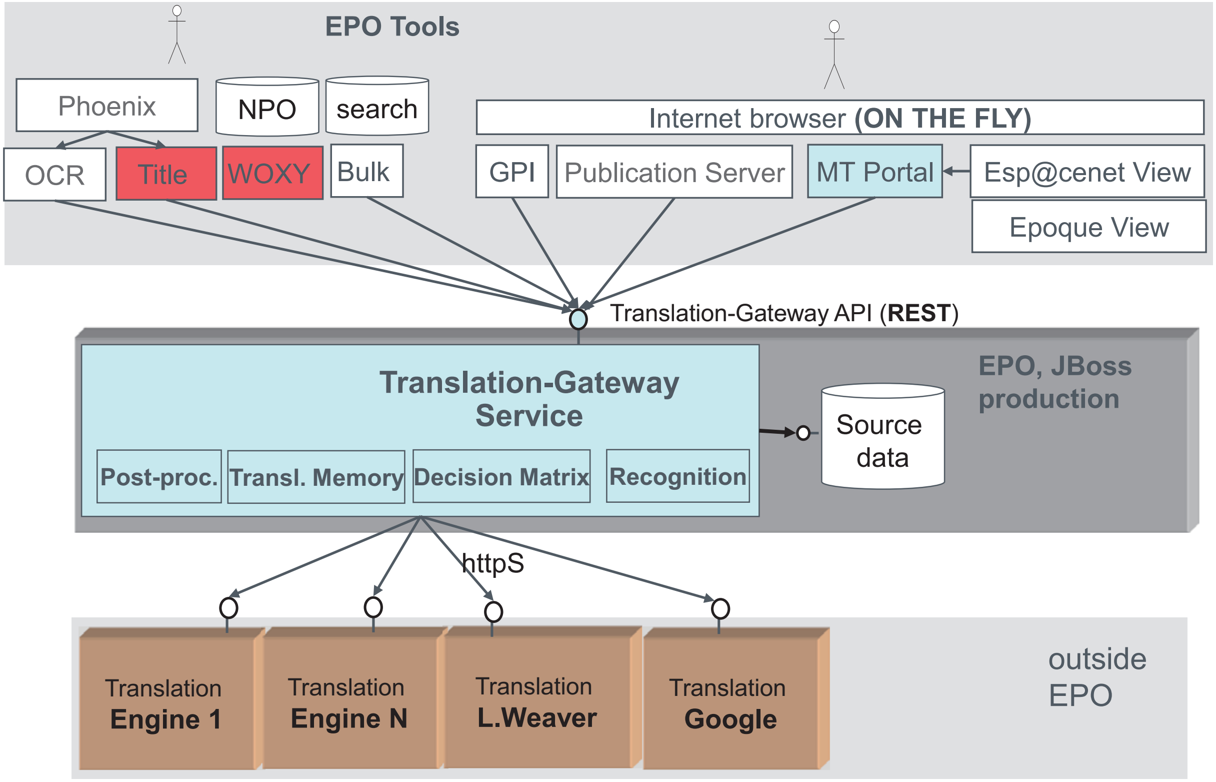
A: Corpora Repository

B: Translation Quality

C. Translation Gateway

D. Engine Pool

Translation Gateway: Service and API



EPO Tools

Phoenix

NPO

search

Internet browser (ON THE FLY)

OCR

Title

WOXY

Bulk

GPI

Publication Server

MT Portal

Esp@cenet View

Epoque View

Translation-Gateway API (**REST**)

Translation-Gateway Service

Post-proc.

Transl. Memory

Decision Matrix

Recognition

Source data

EPO, JBoss production

httpS

Translation Engine 1

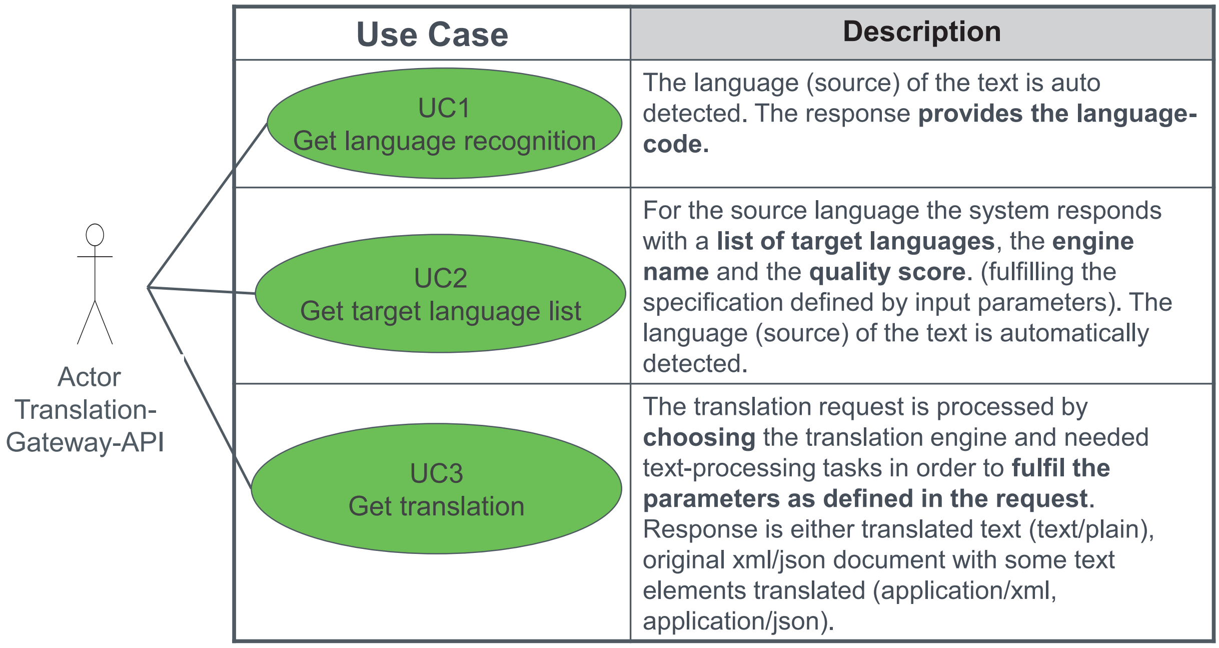
Translation Engine N

Translation L.Weaver

Translation Google

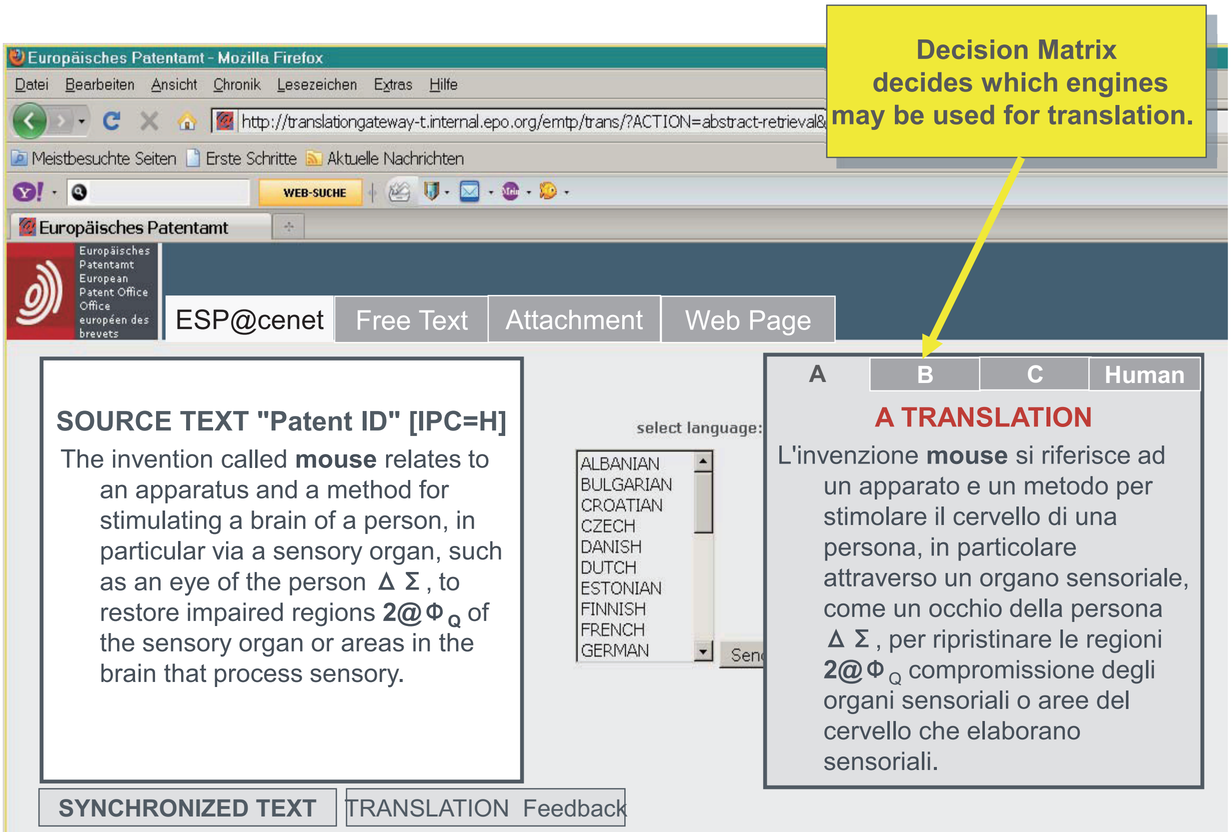
outside EPO

Use Case: Translation-Gateway API



Use Case	Description
UC1 Get language recognition	The language (source) of the text is auto detected. The response **provides the language-code.**
UC2 Get target language list	For the source language the system responds with a **list of target languages**, the **engine name** and the **quality score**. (fulfilling the specification defined by input parameters). The language (source) of the text is automatically detected.
UC3 Get translation	The translation request is processed by **choosing** the translation engine and needed text-processing tasks in order to **fulfil the parameters as defined in the request**. Response is either translated text (text/plain), original xml/json document with some text elements translated (application/xml, application/json).

MTP Portal (concept GUI)



http://translationgateway-t.internal.epo.org/emtp/trans/?ACTION=abstract-retrieval&

Decision Matrix decides which engines may be used for translation.

SOURCE TEXT "Patent ID" [IPC=H]

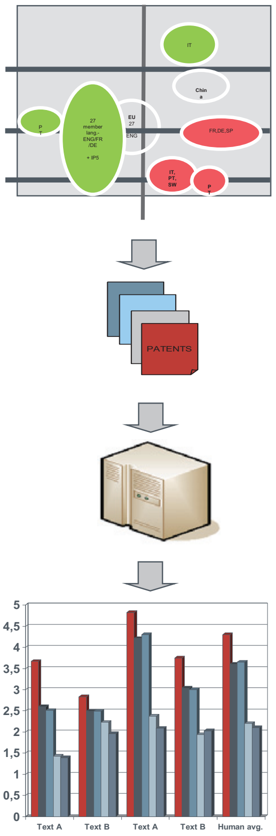
The invention called **mouse** relates to an apparatus and a method for stimulating a brain of a person, in particular via a sensory organ, such as an eye of the person Δ Σ , to restore impaired regions **2@Φ**$_Q$ of the sensory organ or areas in the brain that process sensory.

A TRANSLATION

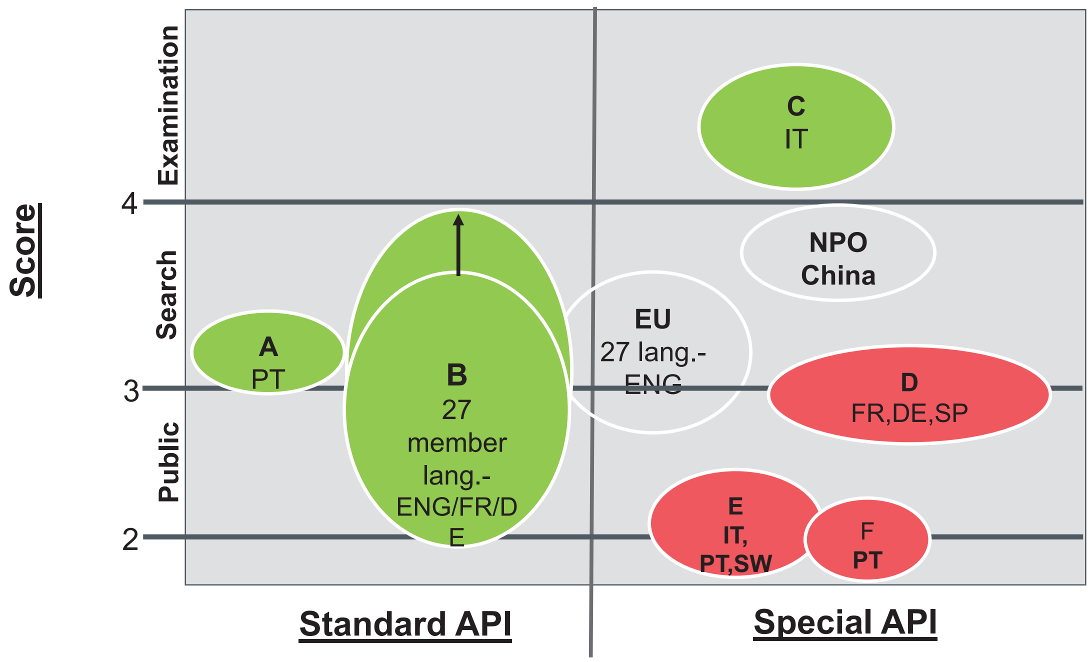
L'invenzione chiamata **mouse** si riferisce ad un apparato e un metodo per stimolare il cervello di una persona, in particolare attraverso un organo sensoriale, come un occhio della persona Δ Σ , per ripristinare le regioni **2@Φ**$_Q$ compromissione degli organi sensoriali o aree del cervello che elaborano sensoriali.

Engine Pool Creation

- **Identify MT engines** that cover the language pair.
  - 30 engines
- **Filter by financial/technical** requirements such as full-text MT solution, server-based, hosted etc.
  - 5 engines
- **Corpora Acquisition:** Human translated documents, needed
- **Training statistics**: Engine provider trains the engine using aligned sentences for translation and documents in source and target language for fluency, quality criteria: 8 weeks
- **Validation:** Acceptance test using native speaking examiners.
  - 1-2 engines



Engine Pool



• Security ISO xxx

Translation On The Fly

Objectives

“Translation on the Fly” will allow examiners to translate any kind of documents with or without OCR with a dynamic selection of the best machine translation for each language and each IPC.

Problem as seen by examiners

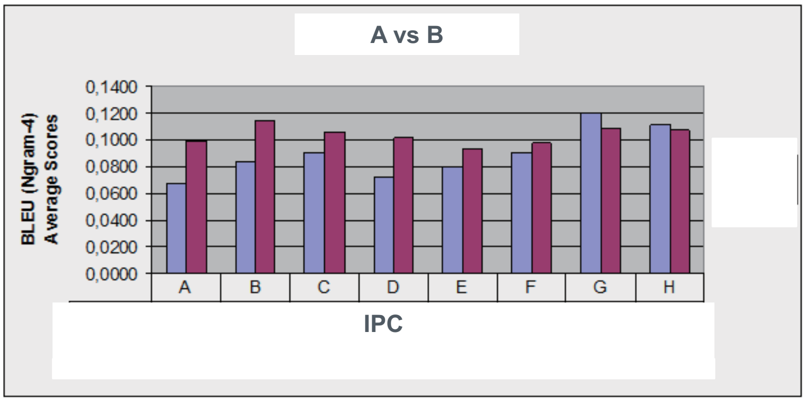
During **search**, **examination** and **classification** Examiners very often need translation of documents not available in English, French or German such as:

- Old or very recent documents (Japanese, PCT, …)
- Russian documents
- Chinese Taipei documents
- Translation of NPL (Chinese NPL in future)
- Translation of internet documents

Solution element 1: Best Translation

⇒ Select best machine translation depending on **language** and **IPC**

– For Italian Language A machine translation could be very good for G&H IPC and B better for the other IPC



Solution element 2: BEST OCR Tool

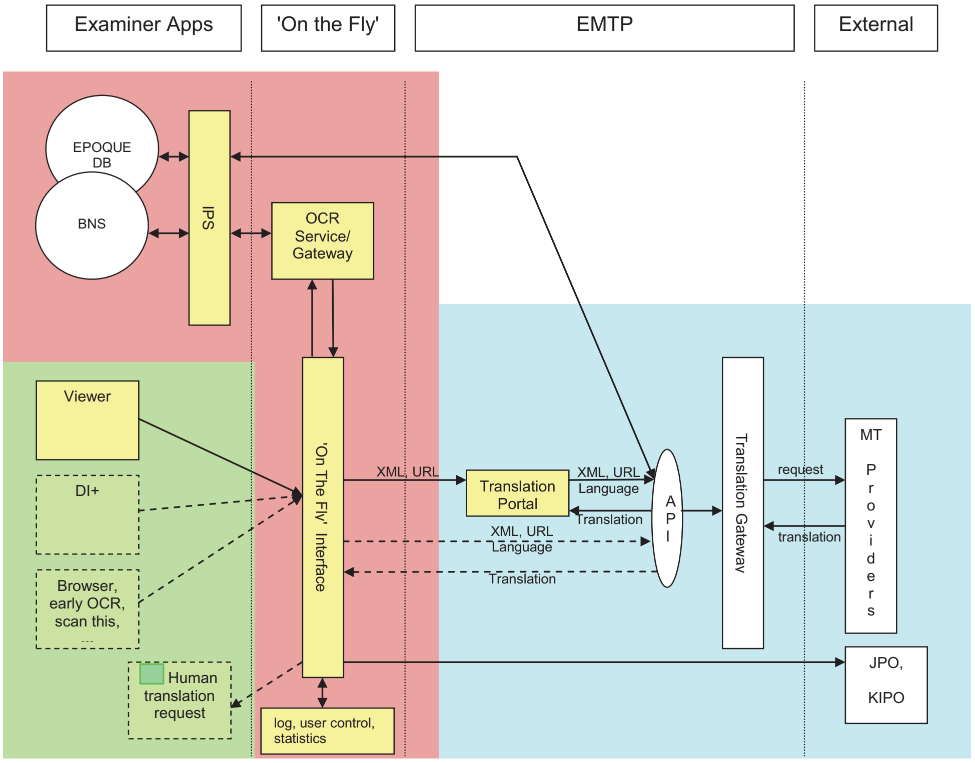
- Select best **quality** OCR tool to access BNS image, especially on Asian languages, Providers might be:

  OMNIPAGE Pro, IRIS, ABBYY

- OCR tool should have a **web API** and must be easily changed if a better OCR tool is on the market (Google OCR)

- OCR tool should be fast and scalable depending on user population

General Architecture



CLIR (query translation)

- **mandatory for patents with no access to Bulk yet**
  Russian, Taiwanese, Swedish, Finish …
- **Non Patent Literature**
- **Price of Bulk translation**
- **Bulk is static, CLIR can be slow**
- **Korean (CLIR), East Meets West discussion**
- **Studies with Bulk vs CLIR**
  - **measure impact of MT quality on different queries**
  - **compare similar queries in BULK MT with Human translation**
  - **compare CLIR engine (EPOQUE with truncation ? +中国专利+** (Zhōngguó zhuānlì)**)**