



# Multilingual text mining and Machine Translation activities carried out at the EC's Joint Research Centre (JRC)

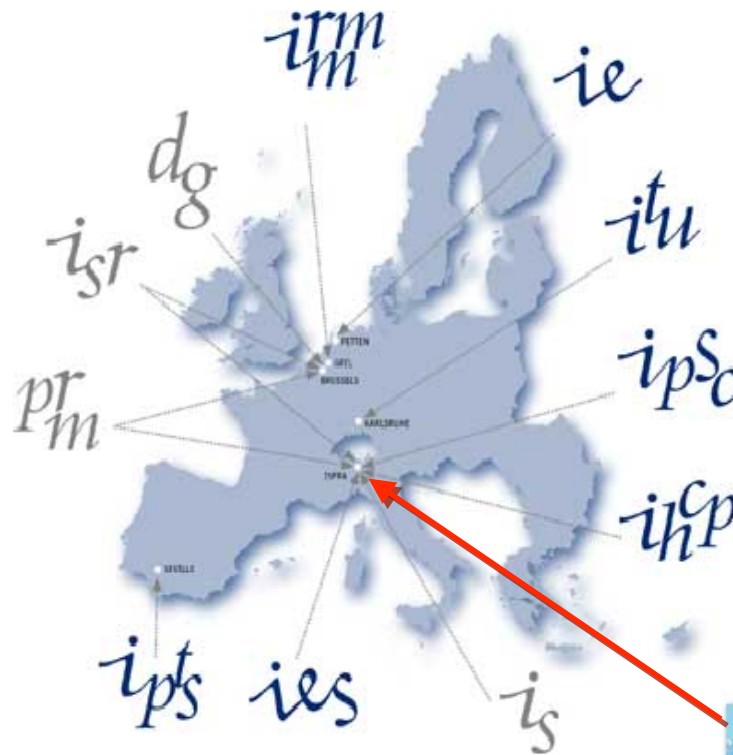


Marco Turchi

& the JRC's *OPTIMA* team – Open Source Text Information Mining and Analysis

*Technical details and publications:* <http://langtech.jrc.ec.europa.eu/>  
*Applications:* <http://emm.newsbrief.eu/overview.html>

- Joint Research Center - Who we are
- Multilinguality:
  - Language-independent algorithm
    - Case Study:  
Multilingual named entity recognition and variant mapping
  - Machine Translation
    - Case Study  
Optima Machine Translation Service



## **BRUSSELS (BE)**

[The Directorate General \(DG\)](#)

[The Institutional and Scientific Relations Directorate \(ISR\)](#)

[The Programme and Resource Management Directorate \(PRM\)](#)

## **GEEL (BE)**

[The Institute for Reference Materials and Measurements \(IRMM\)](#)

## **KARLSRUHE (DE)**

[The Institute for Transuranium Elements \(ITU\)](#)

## **ISPRA (IT)** [Download the Ispra site Brochure \(English - Italian\)](#)

[The Institute for the Protection and Security of the Citizen \(IPSC\)](#)

[The Institute for Environment and Sustainability \(IES\)](#)

[The Institute for Health and Consumer Protection \(IHCP\)](#)

[The Ispra site Directorate \(IS\)](#)

## **PETTEN (NL)**

[The Institute for Energy \(IE\)](#)

## **SEVILLE (E)**

[The Institute for Prospective Technological Studies \(IPT\)](#)



*"The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national."*

# Europe Media Monitor - EMM

**EMM NewsBrief**  
 Explore our projects: EMM NewsBrief | MedSys | EMM Labs  
 Top Stories  
 en - English  
 My Settings

**Main Menu**  
 Top Stories  
 24 Hours Overview  
 Events Detection  
 Most Active Themes  
 Media Edition  
 Help Index EMM  
 Advanced Search  
 Web Site Map

**EU Policy Areas**  
 Agricultural and Media  
 Budget  
 Competition  
 Consumers  
 Culture  
 Customs  
 Development  
 Employment and Monetary Affairs  
 Education, Training, Youth  
 Employment and Social Affairs  
 Energy  
 Environment  
 Enterprise  
 External Relations  
 External Trade  
 Fight against Fraud  
 Fighting Inflation and Financial  
 Food Safety  
 Foreign and Security Policy  
 Human Rights  
 Information Society  
 Internal Market  
 Justice, Freedom and security  
 Public Health  
 Regional Policy  
 Research and Innovation  
 Transport  
 Other

**Current top 10 stories**  
 Languages in Period: Jan 30, 2011 12:10 AM - Jan 30, 2011 12:10 PM  
 Articles per Period: 0h, 1h, 2h, 3h, 4h, 5h, 6h, 7h, 8h, 9h, 10h, 11h, 12h

**Tools**  
 Sunday, January 30, 2011 12:33:00 PM CET  
 RSS XML MAP  
 EMAIL subscribe manage  
 info

**Country Watch**  
 The country most in the news at the moment.  
 Egypt (EG)  
 News about Egypt  
 News sources from Egypt

**MedSys**  
 Updated every 10 minutes, 24 hours per day.  
 all - All languages  
 My Settings

**Most Active Topics**  
 Leprosy  
 In combination with: Spain; India; Brazil;  
 Fontilles, la última leprosería de Europa  
 Miguel ni siquiera se llama Miguel. Hace 60 años dejó atrás su América natal para vivir tras unos muros cuya ubicación y labor ignoran hasta su familia...  
 Fontilles, la última leprosería de Europa  
 start@rds Sunday, January 30, 2011 2:46:00 AM CET | info | en  
 Miguel ni siquiera se llama Miguel. Hace 60 años dejó atrás su América natal para vivir tras unos muros cuya ubicación y labor ignoran hasta su familia...  
 Foot and mouth disease  
 In combination with: Morocco; France; Croatia; Belgium;  
 Au Maroc, le combat inégal de quatre épiéphantèmes face à l'Union européenne  
 LePoint Sunday, January 30, 2011 11:50:00 AM CET | info | en  
 Conçus au Maroc, quatre épiéphantèmes de cirque et leur directeur attendent avec impatience croisées dans un bar de la capitale algérienne.

**Alert level graph**  
 News Items Count  
 Previous 14 days average  
 Alert level: medium low

**EMM BlogBrief**  
 Security  
 Updated every 10 minutes, 24 hours per day.  
 en - English  
 My Settings

**Latest News About - Security**  
 Articles published more than 4 hours ago  
 Conflict intensifies between Histadrut and Government as union battles for higher wages in the public sector  
 Trade Unions Linking Israel and Palestine Trade union leaders from three continents have announced the launch of a new global movement to challenge the obstacles for peace and stability in the area.  
 Comment to: Afghanistan-Taliban Peace Negotiations High-Level Talks With Karzai Government  
 It's possible the US Administration got through to him, convinced him to negotiate some power-sharing with the last-fundamentalist factions and components of the Taliban (which is a generic term, as...  
 Articles published more than 6 hours ago  
 Comment to: Viva Palestine stands with fascist killers of Palestinians  
 "So from '84 to '89 a new War of the Centes began. A war of extermination and extermination, in Shikha and Bouji of Barzagh. The shelling did not stop. Night and day, Shikha was leveled to the ground..."  
 Articles published more than 12 hours ago  
 Comment to: Video Firefighters let home burn to the ground because owner didn't pay annual \$75 fee  
 Anyone else spent hours of us are all really going to agree on the, updated on October 5, 2010 at 5:29 PM No, we won't, but it's enlightening to see who the parts of the good intentions are. ...  
 Articles published more than 12 hours ago  
 David Isenberg The Unaccounted Contractors  
 Author: Shadow Force Private Security Contractors in Iraq Rated: October 5, 2010 03:25 PM Okay, just how long is it going to take for the U.S. government to get an accurate count of the private mil...  
 Comment to: Video Firefighters let home burn to the ground because owner didn't pay annual \$75 fee  
 Okay, so the latest meme is "just charge him fee for services". And if he doesn't pay? Or if he can't pay? Then what? Force him to sell his land? Take it from him? Or do that several of you...  
 Articles published more than 12 hours ago  
 Comment to: Video Firefighters let home burn to the ground because owner didn't pay annual \$75 fee  
 To all the bleeding heart posters in favor of the dp/h/f/h/homewor: Welcome to the Democratic party. Please hand in your guns to the proper governmental official, we will be sure to give them back to...  
 Comment to: Video Firefighters let home burn to the ground because owner didn't pay annual \$75 fee  
 It's quite as long as you promise to (a) stop driving on our roads, (b) stop expressing police protection from crime, (c) stop expressing opinions about what we should do with our army, because it's not...  
 Comment to: Video Firefighters let home burn to the ground because owner didn't pay annual \$75 fee  
 I end up with \$5,000 and a house standing and goodwill in the neighborhood. Actually what you wind up with is a tax lien on a piece of real property with a burned out house on it. You're NEVER going...

**Most reported countries (24h)**  
 Spain: Leprosy (15), India: Leprosy (15), Brazil: Leprosy (16), Morocco: Foot-and-mouth... (6), France: Foot-and-mouth... (6), Croatia: Foot-and-mouth... (6), Belgium: Foot-and-mouth... (6), USA: Endemic (7), France: Endemic (6)

**Daily number of articles in this category**  
 Articles by day: 07/09, 06/10

**NewsDesk Service (a.k.a. RNS) Editorial Interface**

1. Section 1  
 Barroso will Gombauer alle Vice...  
 EU Kommission will sich keinen Mulkorb verpassen lassen...  
 EU Kommission will sich keinen Mulkorb verpassen lassen...  
 EU Presidency and Barroso must react to Berlusconi immediately, says Schulz...  
 IBC: Ireland's reputation needs to be restored...  
 Participation of José Manuel Barroso at the IBC meeting (02/09/2009)...  
 Participation of José Manuel Barroso at the IBC meeting...  
 IBC calls for Yes vote on Lisbon...  
 Barroso Wiederwahl rückt näher...  
 "WIESE" Kommissar von Reinhard Gommel: "Europa-Ehren sind nicht geworden"...  
 Auto-Länder: Balkan-Subventionen sind Senken für die Arbeitsplätze...  
 Barroso attacka commentat l'incertidumbre, però sobre el president...  
 Barroso dirigí EU a blockeare...  
 Putin desobedece al aniversario de la Guerra Mundial exculpando a URSS de sus crímenes...  
 Barroso attacka commentat l'incertidumbre, però sobre el president...  
 Inmortal, para el noticiero de Barroso "El caso é italiano, non siamo irlandesi!"

NewsBrief, Medisys, BlogBrief

EMM Open Source Information Monitoring Engine

- **~ 3,400 Sources** (world-wide, with focus on Europe)
  - news sources (web portals)
  - specialist medical sites
  - commercial newswires
  - 24/7, updated every 10 minutes



- **~ 100,000 articles / day in ~ 50 languages**

- Converts dirty html with adverts, menus, html tags, 'related stories', etc. into clean and standardised UTF-8 encoded RSS format.

- Articles are fed into the various EMM applications/algorithms.



- **~100,000 articles per day... in ~ 50 languages.**
- **Fast!**
- **Research Area:**
  - Multilingual Person Name Recognition
  - Name Variation Matching
  - News Clustering
  - Cross-lingual Linking of News Clusters
  - Document Summarization
  - Sentiment Analysis
  - Automatic Event Extraction
  - Co-reference Resolution
  - Document Classification
  - Machine Translation
  - ...



- **NewsBrief:** current state of affairs, *breaking news* detection in real time
  - **MedSys:** focusing on *health-related* news
  - **NewsExplorer:** long-term, *cross-lingual* news analysis and *people and organization* monitor
  - **EMM-Labs:** various data *visualization* and advanced *text processing* tools
- <http://emm.newsbrief.eu/>

# Why Multilinguality?

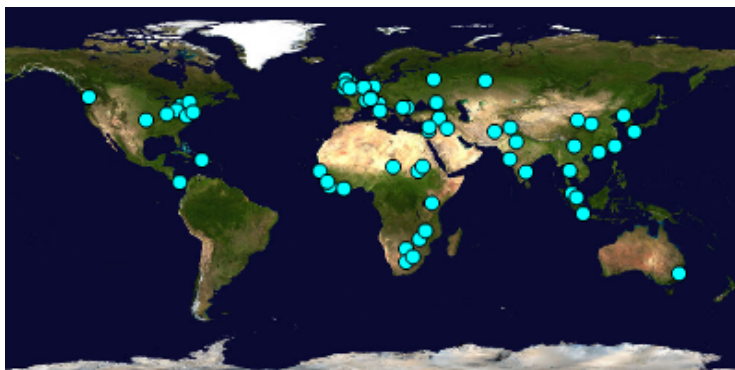
## Locations mentioned in medical articles across languages



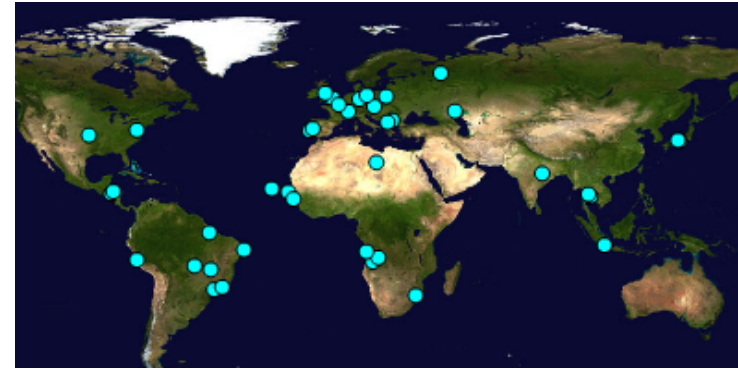
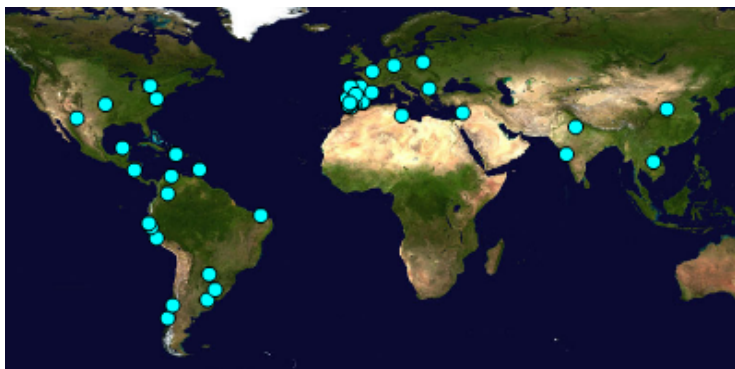
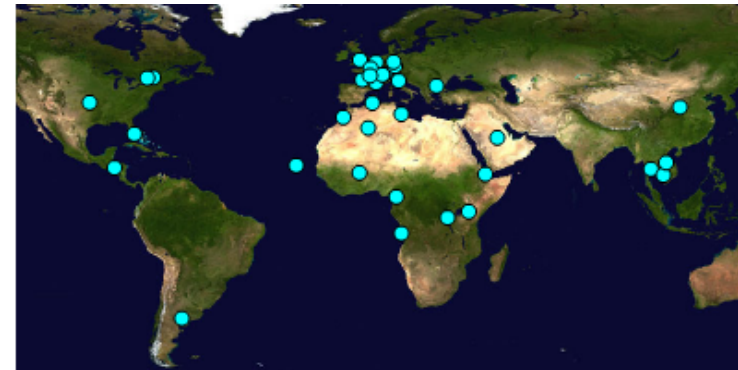
Italian - German



English - French

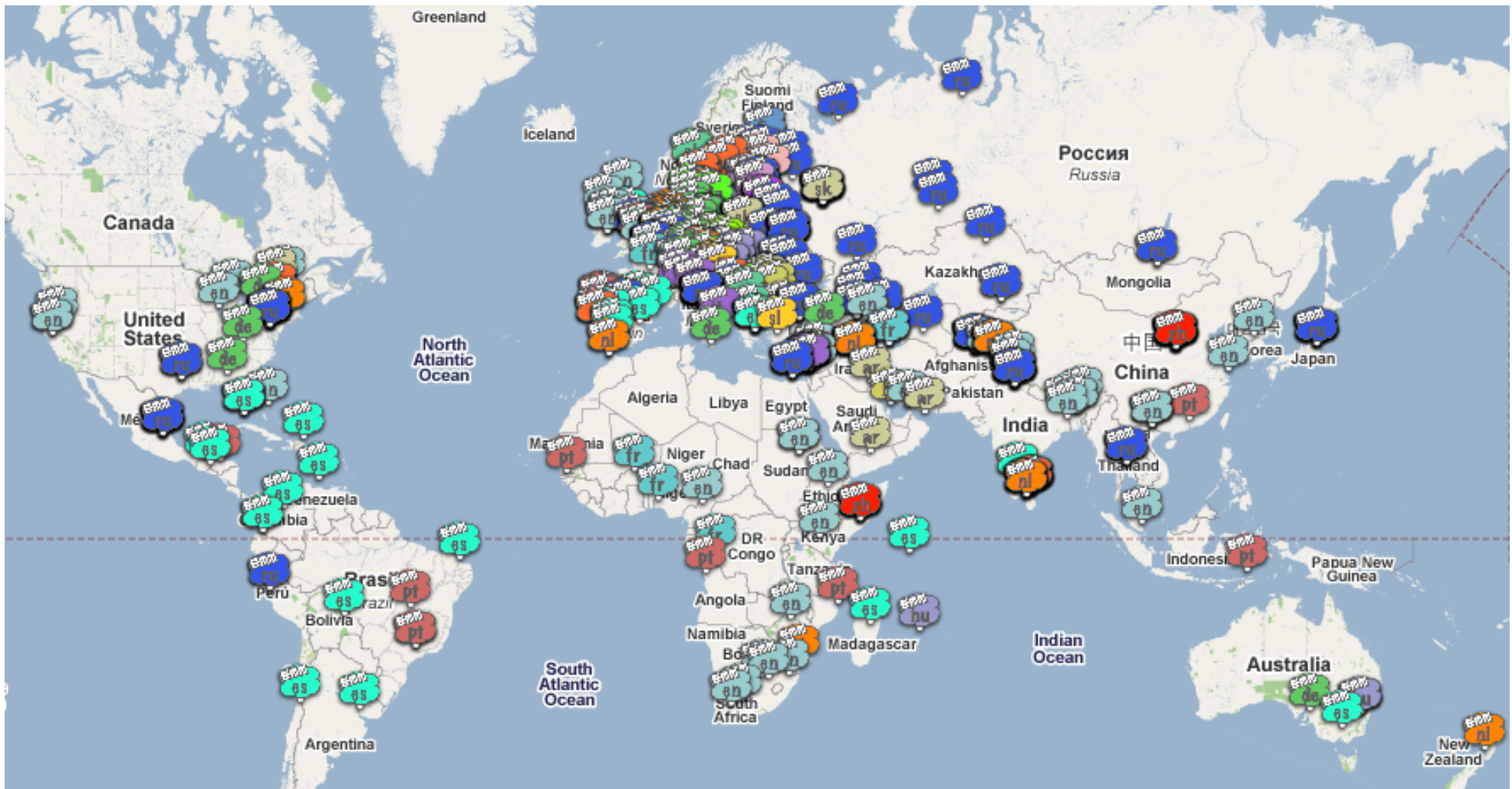


Spanish - Portuguese





Display of latest geo-located news clusters



- **Language-independent** algorithms (as much as possible).
- **Machine Translation :**
  - **Direct:**
    - Translate documents to a common language.
  - **Indirect:**
    - Use the translation engine to improve language independent algorithms.

- Joint Research Center - Who we are
- Multilinguality:
  - Language-independent algorithm
    - Case Study:  
**Multilingual named entity recognition and variant mapping**
  - Machine Translation
    - Case Study  
Optima Machine Translation Service

en death of former Prime Minister **Rafik Hariri**, blamed by many opposition

es asesinato del ex primer ministro **Rafic al-Hariri**, que la oposición atribuyó

fr l'assassinat de l'ex-dirigeant **Rafic Hariri** et le départ du chef de la diplom

nl na de moord op oud-premier **Rafiq al-Hariri** gingen gisteren bijna een

de libanesischen Regierungschef **Rafik Hariri** vor einem Monat wichtige B

sl danjega libanonskega premiera **Rafika Haririja**. Libanonska opozicija si

et möödumisele ekspeaminister **Rafik al-Hariri** surma põhjustanud pommipl

ar اغتيال رئيس الوزراء السابق رفيق الحريري بأياد يهودية وما حدث سابقا

ru Бывший премьер-министр **Ливана Рафик Харири**, который

- **Lookup of known names** from database
  - Currently about 1,150,000 names
  - + 225.000 variants
  - Pre-generate morphological variants (Slovene example):

**Tony**(a|o|u|om|em|m|ju|jem|ja)?\s+**Blair**(a|o|u|om|em|m|ju|jem|ja)

Tony Blair / Tonyju Blairu / Tony Blairt / Tonijs Blērs / Тони Блэра / توني بلير / Tony Blairi / طوني بلير

- **Guessing names** using empirically-derived lexical patterns
  - Trigger word(s) + **Uppercase Words**  
(+ name particles: von, van, de la, abu, bin, ...)
    - **President, Minister, Head of State, Sir, American**
    - “death of”, “[0-9]+-year-old”, ...
    - Combinations: “56-year-old former prime minister **Kurmanbek Bakiyev**”
  - First name lists (**John, Jean, Hans, Giovanni, Johan, ...**)
- Identification of a current average of 608 unknown names per day
- 83 of those are automatically merged with known names
- Use **bootstrapping** to produce a trigger word list for a new language
  - Start with small initial trigger word list or lists of known names
  - Produce frequency list of contexts of known names
  - Manual selection

- Inflection of trigger words for person names, using regular expressions (Slovene example):

- **kandidat**(a|u|om)?
- **legend**(a|e|i|o)
- **milijarder**(ja|ju|jem)?
- **predsednik**(a|u|om|em)?
- **predsednic**(a|e|i|o)
- **ministric**(a|e|i|o)
- **sekretar**(ja|ju|jom|jem)?
- **diktator**(ja|ju|jem)?
- **playboy**(a|u|om|em)?

+ *uppercase words*

... verskega voditelja *Moktade al Sadra* je z notranjim ...  
= *Muqtada al-Sadr* (ID=236 in the DB)



- Adding names (and images) from Wikipedia
- Merging name variants
  - Transliteration
  - Normalisation
  - Similarity measure

in other languages


- Afrikaans
- العربية
- Български
- Dansk
- Deutsch
- Eesti
- Ελληνικά
- Español
- Esperanto
- فارسی
- Français
- Gaeilge
- Galego
- 한국어
- हिन्दी
- Bahasa Indonesia
- Иронау
- Italiano
- עברית
- ಕನ್ನಡ
- Kurdî / كوردی

## Hamid Karzai

From Wikipedia, the free encyclopedia

**Hamid Karzai** (Persian and Pashto: حامد کرزي) (b. December 24, 1957) is the current President of Afghanistan (since December 7, 2004). He became the dominant political figure after the removal of the Taliban government. From 2001, Hamid Karzai was the Chairman of the Transitional Administration and Interim President. In 2004, he won the 2004 election.

**Hamid Karzai**  
حامد کرزي



Хамид Карзай

Hamid Karzai

Hamid Karzaï

Hamid Karsai

حامد کرزاي

हामिद करजई

哈米德·卡尔扎伊



- Currently, EMM NewsExplorer transliterates from Arabic, Farsi, Greek, Russian and Bulgarian
- **Transliterate each character**, or sequence of characters, by a Latin correspondent
  - $\psi \Rightarrow ps$
  - $\lambda \Rightarrow l$
  - $\mu\pi \Rightarrow b$
- Examples of transliterations:
  - Κόφι Ανάν, Greek → ***Kofi Anan***
  - Кофи Аннан, Russian → ***Kofi Annan***
  - Кофи Анан, Bulgarian → ***Kofi Anan***
  - कोफी अन्नान, Hindi → ***Kofi Anan***

- **Transliteration rules** depend on the target language,

e.g.

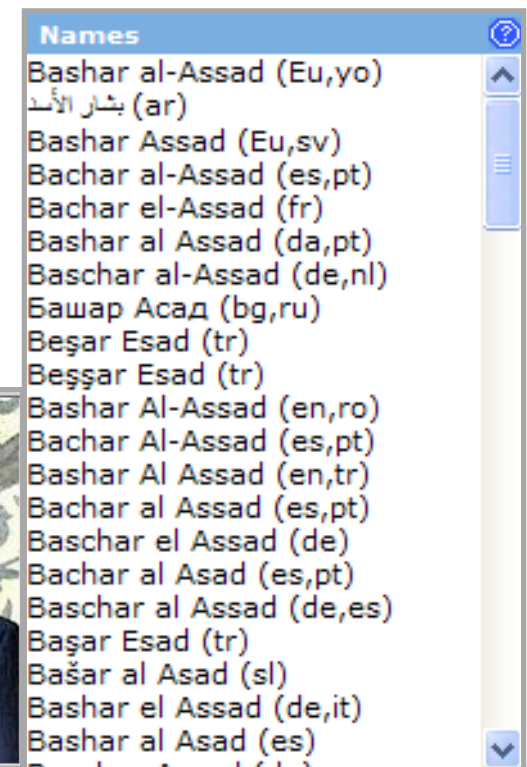
- Владимир Устинов (Russian)
- Vladimir **U**stinov (English)
- **W**ladimir Ustinow (German)
- Vladimir **O**ustinov (French)

- Various ways to represent the **same sound**:

sh, sch, ch, š

e.g.

- Ba**š**ar al Assad
- Ba**sch**ar al Assad
- Ba**ch**ar al Assad



- **Diacritics** are often omitted, e.g.
    - Wałęsa → *Walesa*
    - Saïd → *Said*
    - Schröder → *Schroder*
    - Skarsgård → *Skarsgard*
    - Jørgen → *Jorgen*
- **Edit distance** is large for naturally occurring word variants.

- **Latin normalisation:**

- accented character → non-accented equivalent
- double consonant → single consonant
- ou → u
- “ al-” →
- wl (beginning of name) → vl
- ow (end of name) → ov
- ck → k
- ph → f
- ž → j
- š → sh
- x → ks

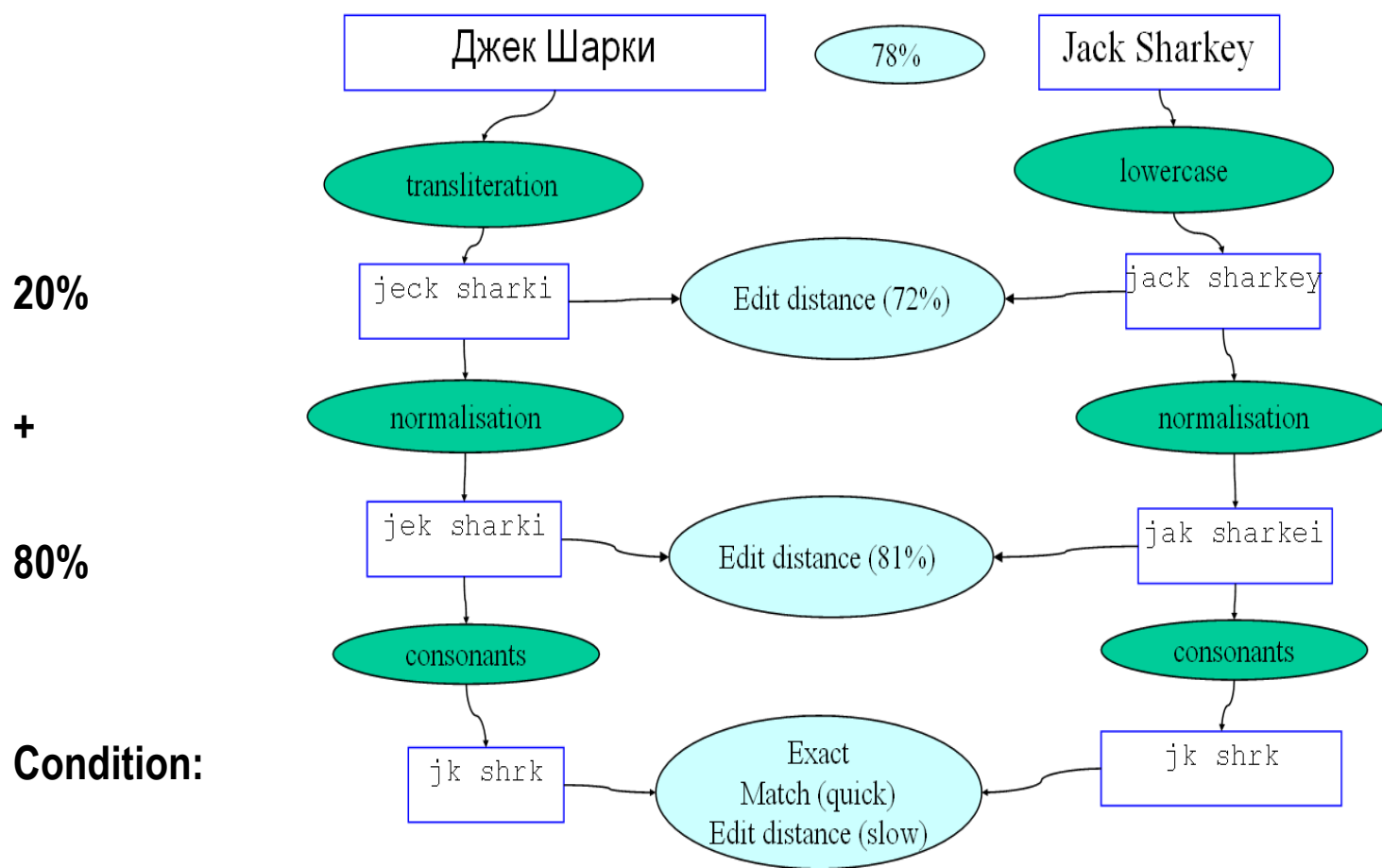
Malik al-Saidoullaiev  
 Malik al-Saidoullaiev  
 Malik al-Saidoulaiev  
 Malik al-Saidulaiev  
 Malik Saidulaiev  
 ... **mlk sdlv**

- **Remove vowels**

Name	Normalised form
Mohammed Siad Barre, Mohamed Siad Barré, Мохаммед Сиад Барре, سياد بري	<b>mhmd sd br</b> (mohamed siad bare)
Mahmoud Ahmadinejad, Mahmūd Ahmadīnežād	<b>mhmd hmdnjd</b> (mahmud ahmadinejad)
Сергей Куприянов, Sergei Kupriyanov, Sergei Kuprianow, Sergueï Kouprianov	<b>srg kprnv</b> (sergei kuprianov)
Ban Ki-moon, Ban Ki Moon, Пан Ги Мун	<b>bn k mn</b> (ban ki mun)

# Similarity measure for name merging

- To compare ~600 new names every day with ~1,375,000 known name variants  
 → **Only if** the transliterated, normalised form with vowels removed is identical
- Calculate edit distance variant similarity using two different representations:



2190. [Muammar Gaddafi](#)

2011-03-24T23:59+0200

Muammar Gaddafi /  
Muammar al-Gaddafi /  
محمّر القذافي / Mouammar  
Kadhafi / Moammar  
Gadhafi / Муамара  
Каддафи / Муаммара  
Каддафи / Muamar  
Gadafi / Muammar  
Gheddafi / 卡扎菲 /  
Muamar al Gadafi /  
Muamar el Gadafi /  
Muammara Kadafiego /  
Muammer Kaddafi /  
Muammar al-Ghadhafi /  
Muammar Kadhafi /  
Муамар Кадафи /  
Muammara Kaddáfi /  
Moamer Kadhafi /  
Muammar al Gaddafi /  
Moamer Gaddafi /

Muammar al Gaddafi /  
Moamer Gaddafi /  
Muammar Gaddafin /  
Moammar Kadhafi /  
Muammar el Gaddafi /  
Muammar Qaddafi /  
Moamerja Gadafija /  
Муамару Каддафи /  
Muammar Kaddafi /  
Muamaro Gaddafi /  
Муаммару Каддафи /  
Muamar Kadafi / Муамар  
Каддафи / Moammar  
Qaddafi / Moamar  
Gadafi / Muammar  
Kaddáfi / Moamer Gadafi /  
Muammar Kadafi / لمحمّر  
القذافي / Muammar Khadafi /  
Muammar Gadhafi /  
Muammar al Gadafi /



Муаммар Каддафи /  
Muammar Gadafi /  
Moammer Kadhafi /  
Muammar el-Qaddafi /  
Muamar Kadhafi /  
Moammar Kadafi / محمّر  
قذافي / Муаммаром  
Каддафи / Moammar  
Gaddafi / Muammar  
Khaddafi / Muamaro  
Kadhafi / Muamar  
Khadafi / Muammar  
Khadafi / Muammer  
Gaddafi / Muamaro  
Gaddafi / Moamar el  
Kadafi / Muhammad  
Gheddafi / Mouammar al-  
Kadhafi / Muammar  
Kadhafi / Muamar El  
Gadafi / Moamar

Gaddafi / Muammar el-  
Gaddafi / Muammar al-  
Gadafi / Mouammar  
Kaddafi / Muammarem  
Kaddáfi / Muammar  
Gadafi / Muammar al-  
Gadhafi / Muammar al-  
Gaddafi / محمّر القذافي /  
Muammar al-Qadhafi /  
Muamarui Gaddafi /  
Muammar Gadaffi /  
Muamaro Kadafio /  
Muamar Gaddafi /  
Muammar Qadhafi /  
Muamaro Gaddafio /  
Muamarui Gaddafiui /  
Muammar al-Gadaffi /  
Muamar al Gaddafi /  
Муаммаром Каддафи /  
Muammar Ghadhafi /

Moamerju Gadafiju /  
Muammarowi  
Kadafiemu / Moammar  
Khadafi / Moammer  
Gaddafi / Muammar El  
Gadafi / מועמר קדאפי /  
Mouamar Kadhafi /  
Mouammar El Kadhafi /  
Muammar Al-Gaddafi /  
Mummar Gaddafi /  
Muamar Gadaffi /  
Moammar Ghadafi

## Name variants

Ali Larijani (Eu,en)  
Ali Laridschani (de)  
Ali Lariyani (es)  
Ali Laridžani (sl)  
علي لاريجاني (ar)  
Alí Larijani (es)  
Ari Larijani (en)  
Alí Lariyani (es)  
Али Лариджани (ru)  
Alì Larijani (it)  
Ali Laridjani (fr)  
Ali Larjani (sv)  
Ali Lariani (it)  
Ali Laryani (es)  
Ali Laranjani (pt)  
Ali Larigani (en)  
Ali Larinjani (nl)  
Al Larijani (nl)  
Ali Ardashir Larijani (en)  
Ali A. Larijani (en)  
Ali Larejani (it)  
Ali Larichani (es)  
علي اردشیر لاریجانی (fa)  
Ali Larijan (en)  
Al Laridschani (de)  
Ali Ardeshir Larijani (en)  
Ali Laridziani (en)



## Trigger words

negotiator (en - 824)  
chefunterhändler (de - 470)  
iraniano (it,pt - 311)  
negociador iraniano (pt - 102)  
iranien (fr - 129)  
iranian negotiator (en - 87)  
pogajalec (sl - 90)  
unterhändler (de - 111)  
iraní (es - 121)  
iraanse onderhandelaar (nl -

A subset of the most important name variants is available:

**[http://  
langtech.jrc.ec.europa.eu/JRC-  
Names.html](http://langtech.jrc.ec.europa.eu/JRC-Names.html)**

- Joint Research Center - Who we are
- Multilinguality:
  - Language-independent algorithm
    - Case Study:  
**Multilingual named entity recognition and variant mapping**
  - Machine Translation
    - Case Study  
**Optima Machine Translation Service**



- **Machine Translation (MT) to directly or indirectly address multilinguality.**
- Two ways of using Statistical Machine Translation:
  - **Direct:**
    - Translate documents to a common language.
  - **Indirect:**
    - Use the translation engine to improve language-independent algorithms.

- EMM gathers around **100k articles** from daily news in **50 languages**.
- Main requirements:
  - Help to **understand the content** of the news articles;
  - **Fast**;
  - **Language-independent**.

- Five key questions when you build a translation service for news and not only:
  1. **Which is the most suitable SMT system for our environment and requirements?**

# Which is the most suitable SMT system?

- Main requirements are **translation speed and quality**
- Options:
  - PBSMT
  - Hierarchical
  - Syntax
- Train and tune the three systems using German-English Europarl V4 and test on news.

System	Time (sec./sent.)	Quality (Bleu Score)
<b>PBSMT</b>	<b>1.02</b>	18.09
<b>Hierarchical</b>	4.5	<b>18.31</b>
<b>Syntax</b>	49	17.62

- Many source languages:
  - **No POS taggers, Syntactic Parsers** (maybe on the target side)
- System combination requires to **translate twice the same sentence.**

- Five key questions when you build a translation service for news and not only :
  1. Which is the most suitable SMT system for our environment?
    - **PBSMT**
  2. **How to translate Named Entities in news?**

- To understand the news content, named entities need to be correctly translated:
  - Source:
    - “*le ministre **bruno le maire** a présenté à **londres** les détails d’ un plan gouvernemental de lutte contre les algues vertes .*”
  - Translated Sentence:
    - “*minister **bruno mayor** has presented in **london** the details of a government plan to fight against the green algae .*”
- Difficult to understand who has presented!!

- **Isolates named entities** (person, location and organization) and **suggests their correct translation** in English to the SMT system:
  - Source:
    - “*le ministre <bruno le maire:bruno le maire> a présenté à <londres:london> les détails d’un plan gouvernemental de lutte contre les algues vertes .*”
  - Translated Sentence:
    - “*minister **bruno le maire** has presented in **london** the details of a government plan to fight against the green algae .*”
- **English translations** are obtained from our database.
- In case of more than one English translations, the most frequent is selected (problem during evaluation: EU or European Union?)

- Five key questions when you build a translation service for news and not only:
  1. Which is the most suitable SMT system for our environment?
    - *PBSMT*
  2. How to translate Named Entities in news?
    - *Suggests their correct translation to the SMT system*
  3. **How to deal with different language styles in news?**



1. *“25 civilians dead as Taliban intensifies attacks in Afghanistan”*
2. *“Twenty-five people were killed in the latest round of Afghan violence this week in Kabul.”*

*Comments?*

- Title and Content have a different language style
  - **Title contains more gerund verbs, no or few linking verbs, prepositions and adverbs.**
    - *“25 civilians dead as Taliban intensifies attacks in Afghanistan”*
  - **Content contains preposition, adverbs, and different verbal tenses.**
    - *“Twenty-five people were killed in the latest round of Afghan violence this week in Kabul.”*
- We trained two instances of the translation system:
  - in the **“title system”**, parameters are optimized using parallel and monolingual titles.
  - in the **“content system”**, parameters are optimized using normal parallel and monolingual sentences.

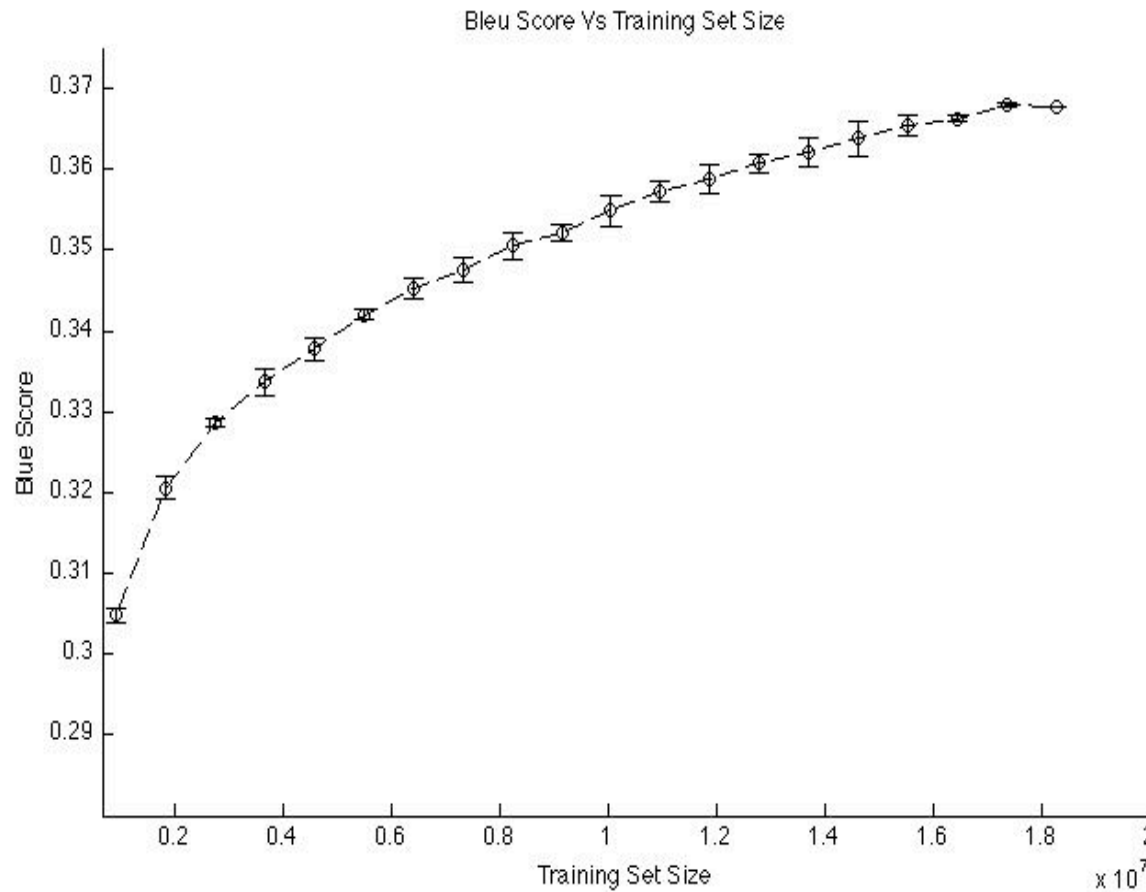
- Some Results for De-En
  - Bleu score

	<b>SMT<sub>Title</sub></b>	<b>SMT<sub>Content</sub></b>
<b>Test-Set<sub>Title</sub></b>	<b>0.3706</b>	0.3511
<b>Test-Set<sub>Content</sub></b>	0.1768	<b>0.1945</b>

- Similar results are obtained for the other language pairs.
- Title data is gathered using Google Translate.

- Five key questions when you build a translation service for news and not only:
  1. Which is the most suitable SMT system for our environment?
    - ***PBSMT***
  2. How to translate Named Entities in news?
    - ***suggests their correct translation to the SMT system***
  3. How to deal with different language styles in news?
    - ***Title and Content customization***
  4. **Which training data can we use?**

# Which training data can we use?

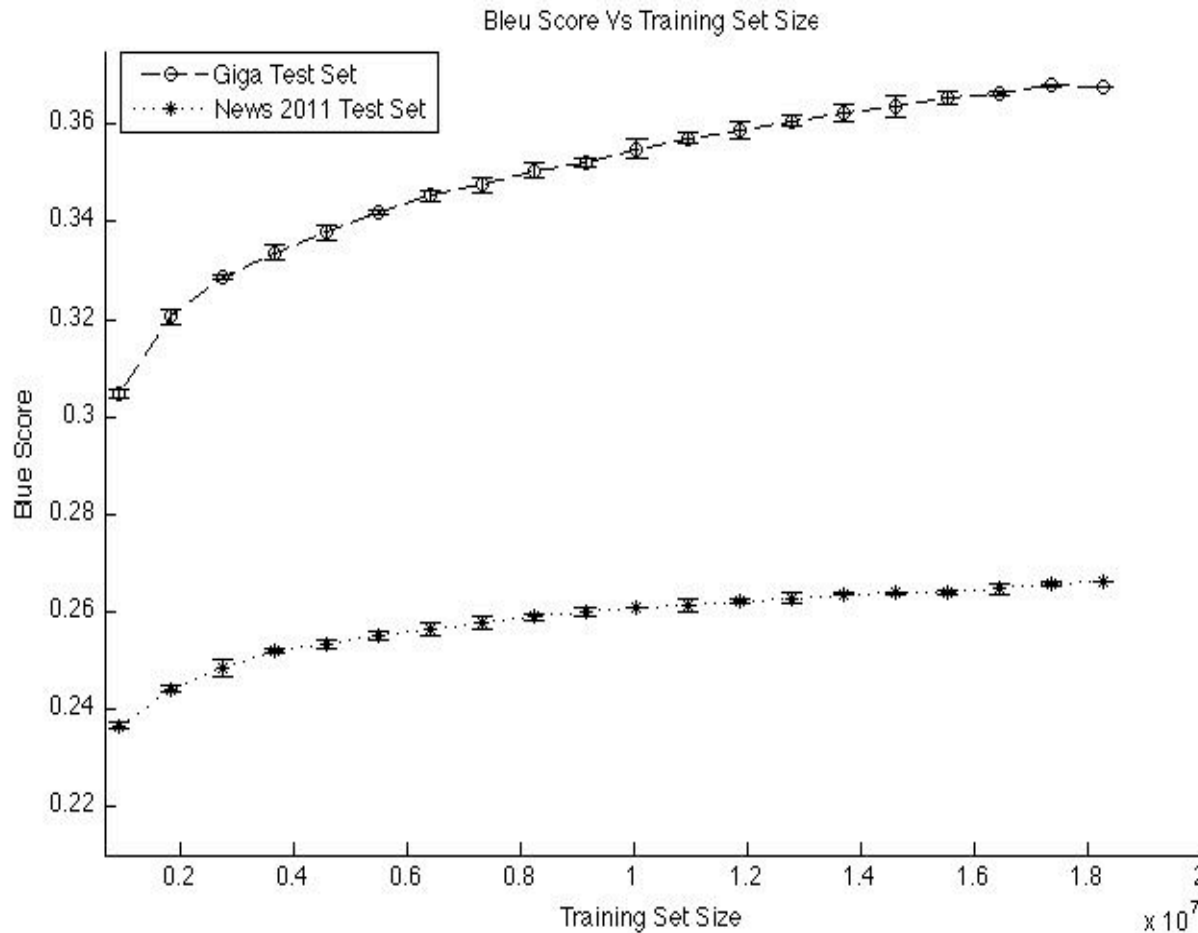


- Learning curve
  - Use Giga corpus (> 18 M sentence pairs) for training and testing.
- 
- Addition of massive amounts of data **from the same distribution** result in small improvements in the performance:
    - same results using different language pairs (**es-en**, **zh-en**) and SMT system (**Moses**, **Portage**).

- Changing test set from the same distribution:
  - Same shape;
  - Small variation in the absolute Bleu score values.
- Considerations:
  - **There are not so huge corpora;**
  - Adding huge amount of data may affect **translation time.**
- **Freely available corpora** (when it is possible):
  - JRC-Acquis, News Commentaries, Europarl, UN Corpora, Movies Subtitle, MultiUN, Opus,...
- Some numbers:
  - Fa ~0.7M
  - De, It, Fr, Es, Pl ~ 4M
  - Ar ~ 6M
  - Tr, Pl > 16M using subtitle corpus.

- Five key questions when you build a translation service for news and not only:
  1. Which is the most suitable SMT system for our environment?
    - ***PBSMT***
  2. How to translate Named Entities in news?
    - ***suggests their correct translation to the SMT system***
  3. How to deal with different language styles in news?
    - ***Title and Content customization***
  4. How many data do we need?
    - ***All freely (and not) available corpora***
  5. **Is the news domain a problem?** (Work in Progress)

# Is the news domain a problem?



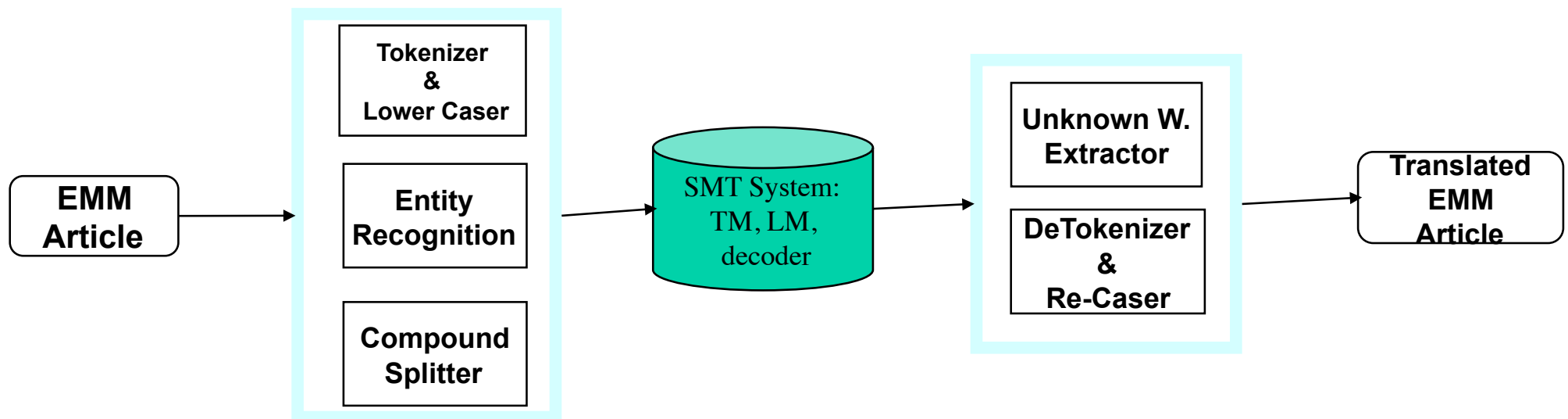
- Testing on news data (out-of-domain):
  - **Drop in performance** compared to the in-domain test;
  - More **flat curve**;
  - **Small gain in Bleu score**.
- **Everyday, news topic changes** according to world events.
  - Always different and new words/phrases to translate.



- News domain is dynamic.
- Parallel corpora
  - Are **static**;
  - Most of them are **not in the news domain**.
- Several approaches have been developed to gather new parallel sentences or fragments.
- Monolingual data varies:
  - daily **update of the language model** using English news of the day (Working in progress).

- Five key questions when you build a translation service for news and not only:
  1. Which is the most suitable SMT system for our environment?
    - ***PBSMT***
  2. How to translate Named Entities in news?
    - ***Suggests their correct translation to the SMT system***
  3. How to deal with different language styles in news?
    - ***Title and Content customization***
  4. Which training data can we use?
    - ***All freely available corpora***
  5. Is domain of the news a problem? (Work in Progress)
    - ***Language model updates***
- **Demo** (Thanks to Česlav Przywara and Yashar Mehdad)

- Translation models available in **Fr, It, Es, De, Fa, Tr, Pl.**
  - **Ar** is coming soon!
- Based on **PBSMT** in **Moses.**
- EMM articles are previously categorized.



URL: Conflicts

Upper Case  Tokenize  Unknown Words  Entities  Compound

## Conflict

**Bülent Ersoy otelden kovuldu** 2011-09-02T11:33+0200

Kıbrıs'ta bir otelde sahne alan Bülent Ersoy, gazetecilere ve seyirciye sataşınca otelden kovuldu.

Time:

Unkown:

## Libya'nın banknotları neden İngiltere'de basıldı ?

**2011-09-02T11:24+0200**

Kaddafi'nin devrilmesi ardından İngiliz uçaklarının ülkeye taşıdığı milyonlarca dinar değerindeki gıcır gıcır banknotun öyküsü.

Time:

**Ayağı kırıldı, 12 saat sonra hastaneye kaldırıldı** 2011-09-02T11:28+0200

Adana'da, yalnız yaşayan 72 yaşındaki yaşlı kadın, gece ayağını kırdı, 12 saat boyunca yardım isteyen kadının çığlıklarını komşuları duydu.

Time:

Unkown:

...Read More

**Çin'de yüzlerce turist ölümden döndü** 2011-09-02T11:28+0200

Çin'in Qiantang nehrinde gel-git izleyenler ölümden döndü. Her sene meydana gelen doğa olayını takip etmek isteyenler, nehrin kenarına toplandığı sırada, nehir sularının altında kaldı.

Time:

Unkown:

...Read More

URL:

Upper Case  Tokenize  Unknown Words  Entities  Compound

## Conflict

**Bulent ersoy hotel fired** 2011-09-02T11:33+0200

A hotel in suva performed bulent ersoy, catpousa, qurrtimulatedannatita hotel and journalists.

## **Libya'nın banknotları neden İngiltere'de basıldı ?** **2011-09-02T11:24+0200**

Kaddafi'nin devrilmesi ardından İngiliz uçaklarının ülkeye taşıdığı milyonlarca dinar değerindeki gıcır gıcır banknotun öyküsü.

Time:

Time: 4.550975

## **Libya banknotes why a f [redacted] eject you know, he could be doing this?** **2011-09-02T11:24+0200**

This british aircraft lovingest Kadhafi toppling the millions of denars worth crisp banknotun story.

Time: 2.1509

[...Read More](#)

**China several hundred turist 's back from the dead** 2011-09-02T11:28+0200

The above qiantang watched in china. Half-girl ormonli follow nine-six fastcar those who, as the river seaside lieth upriver under water.

Time: 2.036529680365297

Unkown:

[...Read More](#)

Communicable Diseases

2011-07-06T10:27+0200 زمان وقوع نوبت عیاشی با نام تنوع

کسبا نوزاد مبتلا به عفونت با نام تنوع (E.COLI) یالک - یا یرتکاب با الکتبا نام نبارد مک نوبش یم مبتلا تنوع بیچوم و نوبش یم یرادا یراچم دراو شیراوگک ملکتسد زا اه نوبرکیم زا یرایسب

Time:  
Unkown:  
[...Read More](#)

2011-06-19T10:42+0200 دراه یویلگ مادمب نرب نراوع ، نامرد زا سب «E.coli» تنوع

عزها با لگ و باسک نبلند لگ اجد باسبا بشو لگ عزها ، ابرها بشو از لگ با دیداد راننده های باسب ، رد تنوع عزها با الکتبا مدم نرب کار طخ و اهدا ب با کسب ، باسبا بشو لگ زا یخرب اما در یگاچم دارا جوی دت با اشورا رد «E.coli» تنوع ، عویش با سب لگ و با حر د : ز و یی کمال س

## چه غذاهائی برای زنان باردار مضر است ؟

زنان باردار باید از مواد غذائی غیر پاستوریزه پرهیز کنند، زیرا موجب آسیب جنین و همچنین مادر میشود و اختلالاتی از جمله آسیب دیدن جفت ، عفونتهای خونی و یا مسموم شدن جنین و یا حتی سقط جنین را منجر میشود

Time:

2011-06-07T10:20+0200

نی لیل با ایجول نراوج ندروخ زا ات داد راننده نادرینش با ناملا لامش رد یلغس سنوبس کاس کلایا ، انس ی شیراز لگ با کسبا ایجول نراوج ی عویش کسبا ناسر کماله با ار رفن 22 نوبت مک «E.coli» عویش اشیم دایز لامتخ با دن یوگیوم ی ناملا کاموم : زوین کمالس دنزوو کننخ دشاب E.coli عویش اشیم نوری لامتخ با مک

Time:  
Unkown:  
[...Read More](#)

2011-07-16T16:06+0200 گنس ارضم رادراب نوز ی ارب ی اذاع هج

ریز ، دننگ زهریب مزو و کسب ریغ ی اذاع داوم زا دیاب رادراب نوز ، دننگیوم دراو ار یکار طخ زوین نوز هج با ، دننگیوم دیدت ار دام کمالس مکن ی ارب موال هج دراد دوچو زوین ی اذاع ، نآ لباوم رد و کسبا دیفم رادراب نوز ی کمالس ی ارب هج دراد دوچو ی اذاع : زوین کمالس .... ی کمال خ و نوبشیم دام نوز چبه و نوزینج بیس ا تخاب

Time:  
Unkown:  
[...Read More](#)

URL:

Upper Case  Tokenize  Unknown Words  Entities  Compound

CommunicableDiseases

Bladder infection is the most infection in women 2011-07-06T10:27+0200

Many **بجورک‌بوم** from digestion device are vital in **پررنا** and bladder infection, including among people in a question - **یرتکاب یالک** (e. coli) the most widespread factor bladder infection in women.

Time: 4.193396226415095

چه غذاهائی برای زنان باردار مضر است ؟

زنان باردار باید از مواد غذائی غیر پاستوریزه پرهیز کنند، زیرا موجب آسیب جنین و همچنین مادر میشود و اختلالاتی از جمله آسیب دیدن جفت ، عفونتهای خونی و یا مسموم شدن جنین و یا حتی سقط جنین را منجر میشود

Time:

Time: 3.554929577464789

Unkown:

## What **UNK** for pregnant women non-target?

Pregnant women should avoid sterile authorisation provisions because cause damage to the fetus and also **UNK** mother and **UNK** including harm to see **UNK UNK** blood or loco fetus or even the eruption of a spontaneous uprising, an abortion.

Time: 2.361

What food for pregnant women non-target? 2011-07-16T16:06+0200

زونی health: there are many food that is good for the health of pregnant women and stand in front of food, there is that in addition to mother **دننگیم** discouraging health threat to a miscarriage also contribute in **دننگیم**. Pregnant women should **نری وکسواب** ingredients or authorisation will not harm because a miscarriage and mother **دوترایم** and **یتال لکخا**...

Time: 2.607669616519174

Unkown:

...Read More

URL: CouncilPresident

 Upper Case  Tokenize  Unknown Words  Entities  Compound  

CouncilPresident

**Rompuy: zmiana konstytucji nie jest niezbędna do stabilizacji finansów** 2011-08-31T21:39+0200

Wpisanie "złotej reguły" budżetowej do konstytucji krajów UE może pomóc w zredukowaniu zadłużenia, ale rządy nie muszą przeprowadzać takiej reformy dla stabilizacji finansów publicznych - oświadczył w środę przewodniczący Rady Europejskiej Herman Van Rompuy.

Time:

Unkown:

## Sikorski o sytuacji w Libii: dziś jest dzień satysfakcji

**2011-08-23T02:41+0200**

Prezydencja Szeft polskiej dyplomacji Radosław Sikorski wyraził nadzieję, że Libijczycy powrócą do społeczności międzynarodowej, z demokratycznie wybranymi władzami. W czasie poniedziałkowego briefingu poświęconego sytuacji w Libii, nie chciał natomiast komentować doniesień o tym, że Polska dostarczała broń libijskim powstańcom.

Time:

Unkown:

[...Read More](#)**Sikorski o sytuacji w Libii: dziś jest dzień satysfakcji** 2011-08-23T02:41+0200

Prezydencja Szeft polskiej dyplomacji Radosław Sikorski wyraził nadzieję, że Libijczycy powrócą do społeczności międzynarodowej, z demokratycznie wybranymi władzami. W czasie poniedziałkowego briefingu poświęconego sytuacji w Libii, nie chciał natomiast komentować doniesień o tym, że Polska dostarczała broń libijskim powstańcom.

Time:

Unkown:

[...Read More](#)



URL: CouncilPresident

Upper Case  Tokenize  Unknown Words  Entities  Compound  pl

## Sikorski o sytuacji w Libii: dziś jest dzień satysfakcji

2011-08-23T02:41+0200

Prezydencja Szef polskiej dyplomacji Radosław Sikorski wyraził nadzieję, że Libijczycy powrócą do społeczności międzynarodowej, z demokratycznie wybranymi władzami. W czasie poniedziałkowego briefingu poświęconego sytuacji w Libii, nie chciał natomiast komentować doniesień o tym, że Polska dostarczała broń libijskim powstańcom.

Time:

[Read More](#)

## Sikorski on the situation in Libya: today is the day of satisfaction

2011-08-23T02:41+0200

The head of the Polish diplomacy Radoslaw Sikorski expressed the hope that *libijczycy* will return to the international community of democratically elected authorities. On Monday during a briefing on the situation in Libya, however, did not want to comment on reports that Poland *powstańcom* Libyan supplied weapons.

Time: 1.55

Time: 1.5584415584415585

Unkown:

...[Read More](#)

- **Word ordering** in the target sentence.
- **Vocabulary coverage** in agglutinative language, e.g. Tr.
- **Select the most reliable translated articles** to show to the final user
  - Confidence estimation approaches.
- **Training data for less resourced languages** or approaches to **expand translation capability without parallel data**
  - Lithuanian, Farsi, Croatian, Swahili,...
- **Subtitle corpora**
  - Pl-En 4M training sentences without subtitles: 57.84 Bleu score;
  - Pl-En 19M training sentences with subtitles: 57.36 Bleu score;
  - Tr-En with subtitle: high frequency of slang words.
- Different perception of translation quality (ready to luggage pack)
- ...

- Large amount of data in more than 50 languages are gathered by the EMM's family every day.
- Language-independent algorithms:
  - Quite a **successful story at JRC**:
    - Multilingual named entity recognition and variant mapping
    - ...
- Machine Translation is a **new research area at the JRC**.
  - Translation service based on Moses:
    - Translate documents to a common language.
  - Use the translation engine to improve language independent algorithms.



# Multilingual text mining and Machine Translation activities carried out at the EC's Joint Research Centre (JRC)



Marco Turchi

& the JRC's *OPTIMA* team – Open Source Text Information Mining and Analysis

*Technical details and publications:* <http://langtech.jrc.ec.europa.eu/>  
*Applications:* <http://emm.newsbrief.eu/overview.html>