Cross-lingual similarity calculation for plagiarism detection and more – Tools and resources

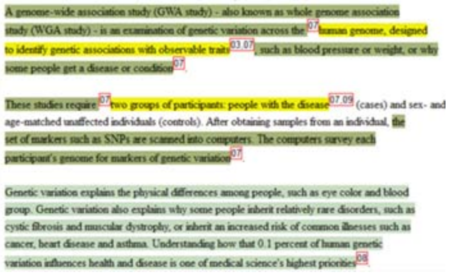
Ralf Steinberger

European Commission – Joint Research Centre (JRC)
http://langtech.jrc.ec.europa.eu/

PAN
PAN 2012 Lab
Uncovering Plagiarism, Authorship, and Social Software Misuse
held in conjunction with the CLEF 2012 Conference
17-20 September 2012, Rome, Italy
ROME 2012
Clef

Agenda

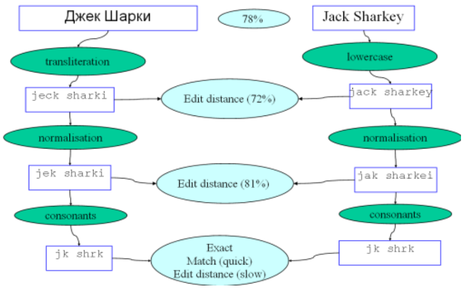
- EC-Joint Research Centre (JRC) – Who we are
- Monolingual plagiarism detection (PD) work at the JRC
- Cross-lingual similarity calculation at the JRC
    - Named entity (NE) matching across languages
    - Linking related news items across languages
    - Identifying translations of documents
- JRC's multilingual tools and resources
- Summary

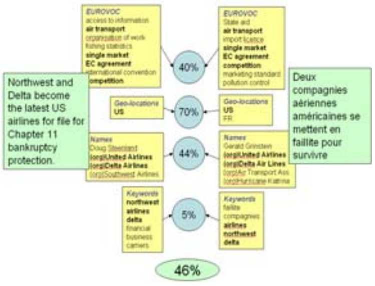


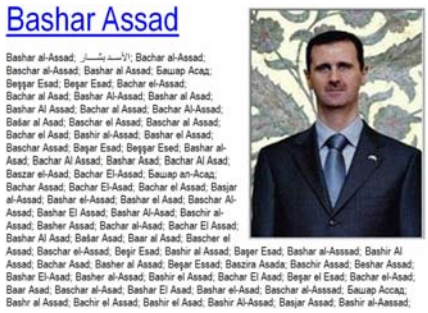
A genome-wide association study (GWA study) - also known as whole genome association study (WGA study) - is an examination of genetic variation across the [07] human genome, designed to identify genetic associations with observable traits [03.07], such as blood pressure or weight, or why some people get a disease or condition [07].

These studies require [07] two groups of participants: people with the disease [07.09] (cases) and sex- and age-matched unaffected individuals (controls). After obtaining samples from an individual, the set of markers such as SNPs are scanned into computers. The computers survey each participant's genome for markers of genetic variation [07].

Genetic variation explains the physical differences among people, such as eye color and blood group. Genetic variation also explains why some people inherit relatively rare disorders, such as cystic fibrosis and muscular dystrophy, or inherit an increased risk of common illnesses such as cancer, heart disease and asthma. Understanding how that 0.1 percent of human genetic variation influences health and disease is one of medical science's highest priorities [08].



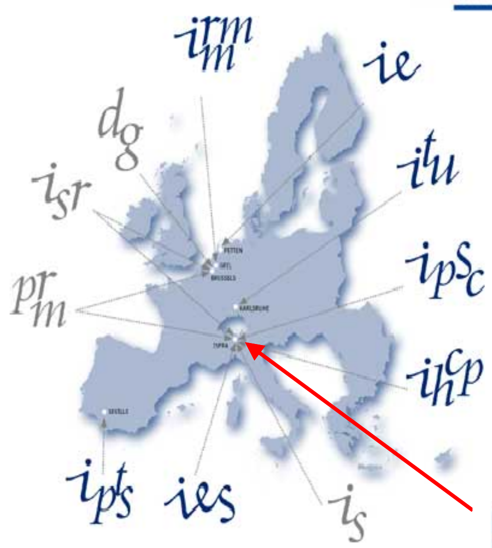




Bashar Assad

Bashar al-Assad; الأسد بشار; Bachar al-Assad; Baschar al-Assad; Bashar al Assad; Башар Асад; Beşşar Esad; Beşar Esad; Bachar el-Assad; Bachar al Asad; Bashar Al-Assad; Bashar al Asad; Bashar Al Assad; Bachar al Assad; Bachar Al-Assad; Bašar al Asad; Baschar el Assad; Baschar al Assad; Bachar el Asad; Bashir al-Assad; Bashar el Assad; Baschar Assad; Başar Esad; Beşşar Esed; Bashar al-Asad; Bachar Al Assad; Bashar Asad; Bachar Al Asad; Baszar el-Asad; Bachar El-Assad; Башар ал-Асад; Bachar Assad; Bachar El-Asad; Bachar el Assad; Basjar al-Assad; Bashar el-Assad; Bashar el Asad; Baschar Al-Assad; Bashar El Assad; Bashar Al-Asad; Baschir al-Assad; Basher Assad; Bachar al-Asad; Bachar El Assad; Bashar Al Asad; Bašar Asad; Baar al Asad; Bascher el Assad; Baschar el-Assad; Beşir Esad; Bashir al Assad; Başer Esad; Bashar al-Asssad; Bashir Al Assad; Bachar Asad; Basher al Assad; Beşar Essad; Baszira Asada; Baschir Assad; Beshar Assad; Bashar El-Asad; Basher al-Assad; Bashir el Assad; Bachar El Asad; Beşar el Esad; Bachar el-Asad; Baar Asad; Baschar al-Asad; Bashar El Asad; Bashar el-Asad; Baschar al-Asssad; Башар Ассад; Bashr al Assad; Bachir el Assad; Bashir el Asad; Bashir Al-Assad; Basjar Assad; Bashir al-Aassad;

JRC - Who we are



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 (**IPTS**)

- European Commission
  (scientific-technical arm of public administration)
- Non-commercial
- Multi-disciplinary / multilingual
- Main product: Europe Media Monitor (EMM)

Europe Media Monitor

EMM – A few facts

- ~ 150,000 online news articles / day in ~ 50 languages
- ~ 3600 Sources (world-wide, with focus on Europe)

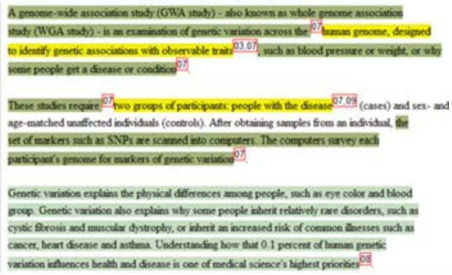
- In-depth analysis in 20 languages (NewsExplorer)
- 24/7, updated every 10 minutes
- Freely accessible via http://emm.newsbrief.eu/overview.html

- Articles are fed into the various EMM applications:

Steinberger Ralf, Bruno Pouliquen & Erik van der Goot (2009). An introduction to the Europe Media Monitor Family of Applications. In: Fredric Gey, Noriko Kando & Jussi Karlgren (eds.): Information Access in a Multilingual World - Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009), pp. 1-8. Boston, USA. 23 July 2009.

Agenda

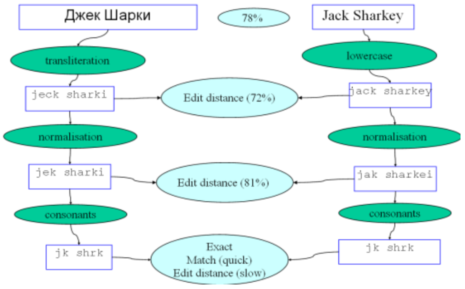
- EC-Joint Research Centre (JRC) – Who we are
- Monolingual plagiarism detection (PD) work at the JRC
- Cross-lingual similarity calculation at the JRC
  - Named entity (NE) matching across languages
  - Linking related news items across languages
  - Identifying translations of documents
- JRC's multilingual tools and resources
- Summary

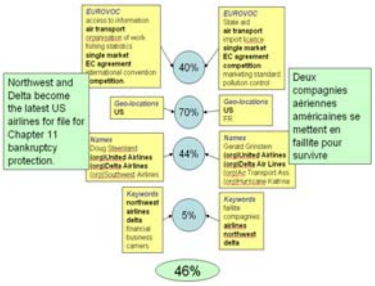


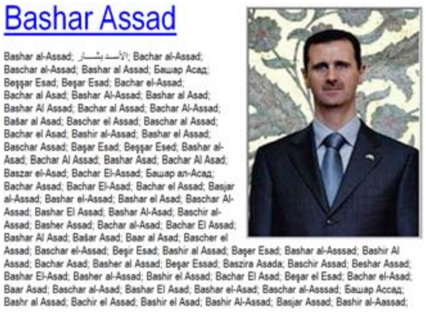
A genome-wide association study (GWA study) - also known as whole genome association study (WGA study) - is an examination of genetic variation across the [07] human genome, designed to identify genetic associations with observable traits [03.07], such as blood pressure or weight, or why some people get a disease or condition [07].

These studies require [07] two groups of participants: people with the disease [07.09] (cases) and sex- and age-matched unaffected individuals (controls). After obtaining samples from an individual, the set of markers such as SNPs are scanned into computers. The computers survey each participant's genome for markers of genetic variation [07].

Genetic variation explains the physical differences among people, such as eye color and blood group. Genetic variation also explains why some people inherit relatively rare disorders, such as cystic fibrosis and muscular dystrophy, or inherit an increased risk of common illnesses such as cancer, heart disease and asthma. Understanding how that 0.1 percent of human genetic variation influences health and disease is one of medical science's highest priorities [08].





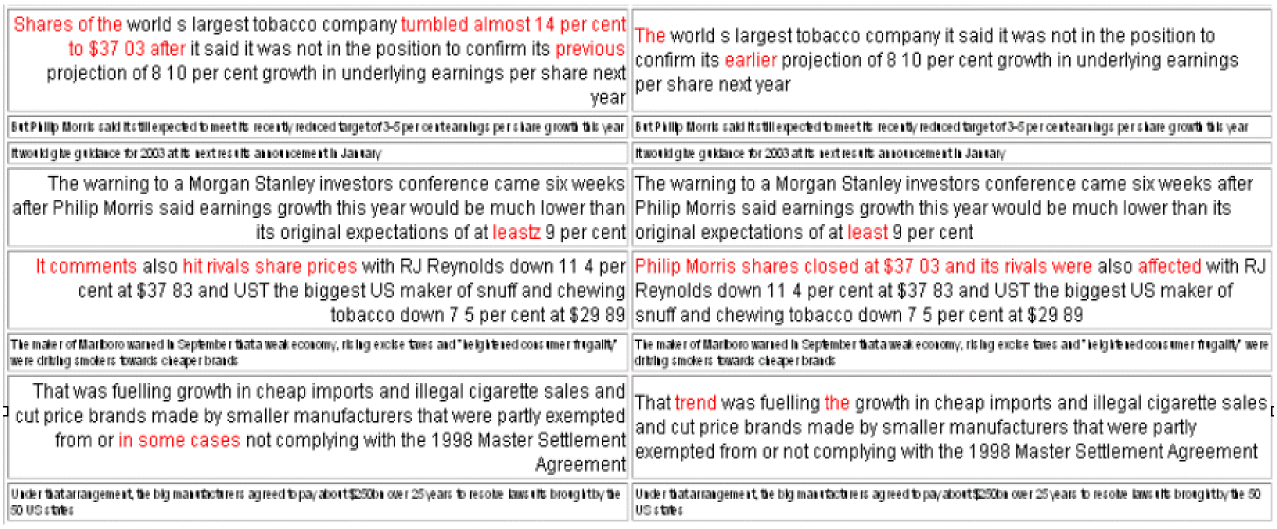


Bashar Assad

Bashar al-Assad; الأسد بشار; Bachar al-Assad; Baschar al-Assad; Bashar al Assad; Башар Асад; Beşşar Esad; Beşar Esad; Bachar el-Assad; Bachar al Asad; Bashar Al-Assad; Bashar al Asad; Bashar Al Assad; Bachar al Assad; Bachar Al-Assad; Bašar al Asad; Baschar el Assad; Baschar al Assad; Bachar el Asad; Bashir al-Assad; Bashar el Assad; Baschar Assad; Başar Esad; Beşşar Esed; Bashar al-Asad; Bachar Al Assad; Bashar Asad; Bachar Al Asad; Baszar el-Asad; Bachar El-Assad; Башар ал-Асад; Bachar Assad; Bachar El-Asad; Bachar el Assad; Basjar al-Assad; Bashar el-Assad; Bashar el Asad; Baschar Al-Assad; Bashar El Assad; Bashar Al-Asad; Baschir al-Assad; Basher Assad; Bachar al-Asad; Bachar El Assad; Bashar Al Asad; Bašar Asad; Baar al Asad; Bascher el Assad; Baschar el-Assad; Beşir Esad; Bashir al Assad; Başer Esad; Bashar al-Asssad; Bashir Al Assad; Bachar Asad; Basher al Assad; Beşar Essad; Baszira Asada; Baschir Assad; Beshar Assad; Bashar El-Asad; Basher al-Assad; Bashir el Assad; Bachar El Asad; Beşar el Esad; Bachar el-Asad; Baar Asad; Baschar al-Asad; Bashar El Asad; Bashar el-Asad; Baschar al-Asssad; Башар Ассад; Bashr al Assad; Bachir el Assad; Bashir el Asad; Bashir Al-Assad; Basjar Assad; Bashir al-Aassad;

Monolingual PD work at the JRC (1)

- N-gram overlap between pairs of documents
  - Karp-Rabin algorithm, using word 5-grams
  - to weed out duplicates in the IAEA document database (ca. 350K documents)
  - to find news article near-duplicates in EMM (applied to all news clusters)



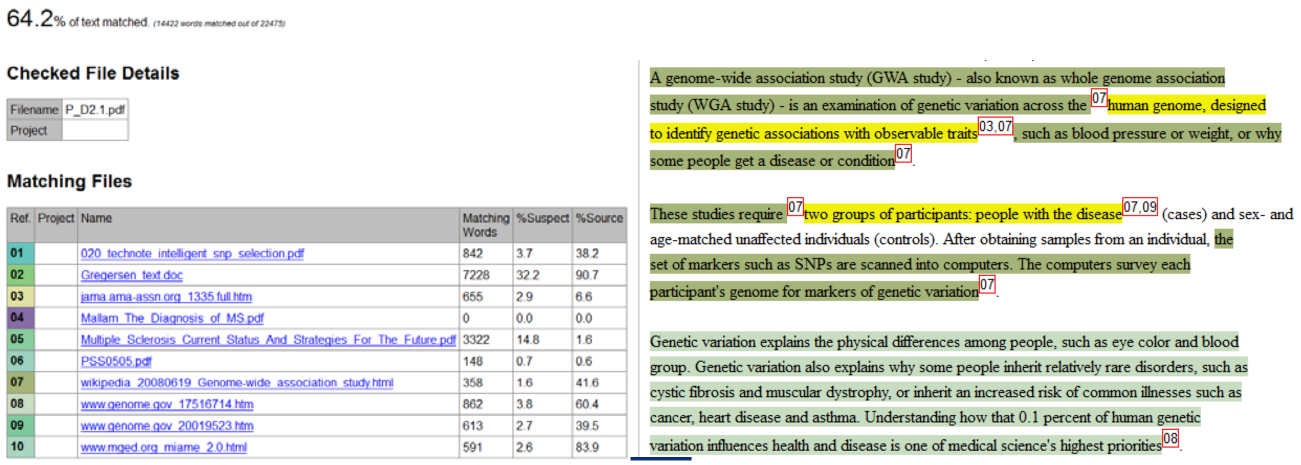
Shares of the world s largest tobacco company tumbled almost 14 per cent to $37 03 after it said it was not in the position to confirm its previous projection of 8 10 per cent growth in underlying earnings per share next year	The world s largest tobacco company it said it was not in the position to confirm its earlier projection of 8 10 per cent growth in underlying earnings per share next year
But Philip Morris said it still expected to meet its recently reduced target of 3-5 per cent earnings per share growth this year	But Philip Morris said it still expected to meet its recently reduced target of 3-5 per cent earnings per share growth this year
It would give guidance for 2003 at its next results announcement in January	It would give guidance for 2003 at its next results announcement in January
The warning to a Morgan Stanley investors conference came six weeks after Philip Morris said earnings growth this year would be much lower than its original expectations of at leastz 9 per cent	The warning to a Morgan Stanley investors conference came six weeks after Philip Morris said earnings growth this year would be much lower than its original expectations of at least 9 per cent
It comments also hit rivals share prices with RJ Reynolds down 11 4 per cent at $37 83 and UST the biggest US maker of snuff and chewing tobacco down 7 5 per cent at $29 89	Philip Morris shares closed at $37 03 and its rivals were also affected with RJ Reynolds down 11 4 per cent at $37 83 and UST the biggest US maker of snuff and chewing tobacco down 7 5 per cent at $29 89
The maker of Marlboro warned in September that a weak economy, rising excise taxes and "heightened consumer frugality" were driving smokers towards cheaper brands	The maker of Marlboro warned in September that a weak economy, rising excise taxes and "heightened consumer frugality" were driving smokers towards cheaper brands
That was fuelling growth in cheap imports and illegal cigarette sales and cut price brands made by smaller manufacturers that were partly exempted from or in some cases not complying with the 1998 Master Settlement Agreement	That trend was fuelling the growth in cheap imports and illegal cigarette sales and cut price brands made by smaller manufacturers that were partly exempted from or not complying with the 1998 Master Settlement Agreement
Under that arrangement, the big manufacturers agreed to pay about $250bn over 25 years to resolve lawsuits brought by the 50 US states	Under that arrangement, the big manufacturers agreed to pay about $250bn over 25 years to resolve lawsuits brought by the 50 US states

Monolingual PD work at the JRC (2)

European Commission

Detection of verbatim plagiarism in research deliverables of EC-funded projects.

- Method: Search for longest (in chars) word 6-grams of each document in EC database and on the web (avoiding strings from document template)
  - If target documents pass similarity threshold:
  - Full-text comparison of matching documents to detect significant matches
  - Visualise document overlap and manually check.
- Contact: Charles Macmillan



64.2% of text matched. (14422 words matched out of 22475)

Checked File Details

Filename	P_D2.1.pdf
Project	

Matching Files

Ref.	Project	Name	Matching Words	%Suspect	%Source
01		020_technote_intelligent_snp_selection.pdf	842	3.7	38.2
02		Gregersen_text.doc	7228	32.2	90.7
03		jama.ama-assn.org_1335.full.htm	655	2.9	6.6
04		Mallam_The_Diagnosis_of_MS.pdf	0	0.0	0.0
05		Multiple_Sclerosis_Current_Status_And_Strategies_For_The_Future.pdf	3322	14.8	1.6
06		PSS0505.pdf	148	0.7	0.6
07		wikipedia_20080619_Genome-wide_association_study.html	358	1.6	41.6
08		www.genome.gov_17516714.htm	862	3.8	60.4
09		www.genome.gov_20019523.htm	613	2.7	39.5
10		www.mged.org_miame_2.0.html	591	2.6	83.9

A genome-wide association study (GWA study) - also known as whole genome association study (WGA study) - is an examination of genetic variation across the [07] human genome, designed to identify genetic associations with observable traits [03,07], such as blood pressure or weight, or why some people get a disease or condition [07].

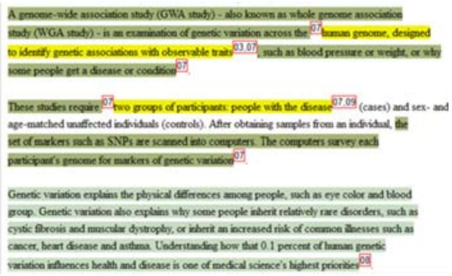
These studies require [07] two groups of participants: people with the disease [07,09] (cases) and sex- and age-matched unaffected individuals (controls). After obtaining samples from an individual, the set of markers such as SNPs are scanned into computers. The computers survey each participant's genome for markers of genetic variation [07].

Genetic variation explains the physical differences among people, such as eye color and blood group. Genetic variation also explains why some people inherit relatively rare disorders, such as cystic fibrosis and muscular dystrophy, or inherit an increased risk of common illnesses such as cancer, heart disease and asthma. Understanding how that 0.1 percent of human genetic variation influences health and disease is one of medical science's highest priorities [08].

Agenda

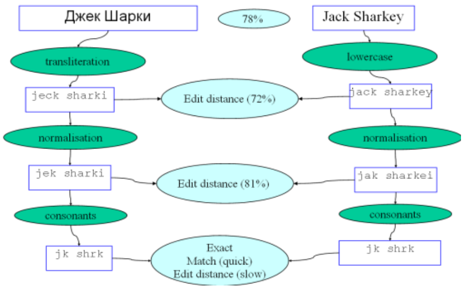
European Commission

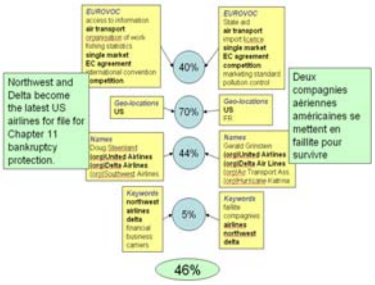
- EC-Joint Research Centre (JRC) – Who we are
- Monolingual plagiarism detection (PD) work at the JRC
- Cross-lingual similarity calculation at the JRC
  - Named entity (NE) matching across languages
  - Linking related news items across languages
  - Identifying translations of documents
- JRC's multilingual tools and resources
- Summary


A genome-wide association study (GWA study) - also known as whole genome association study (WGA study) - is an examination of genetic variation across the [07] human genome, designed to identify genetic associations with observable traits [03,07], such as blood pressure or weight, or why some people get a disease or condition [07].

These studies require [07] two groups of participants: people with the disease [07,09] (cases) and sex- and age-matched unaffected individuals (controls). After obtaining samples from an individual, the set of markers such as SNPs are scanned into computers. The computers survey each participant's genome for markers of genetic variation [07].

Genetic variation explains the physical differences among people, such as eye color and blood group. Genetic variation also explains why some people inherit relatively rare disorders, such as cystic fibrosis and muscular dystrophy, or inherit an increased risk of common illnesses such as cancer, heart disease and asthma. Understanding how that 0.1 percent of human genetic variation influences health and disease is one of medical science's highest priorities [08].

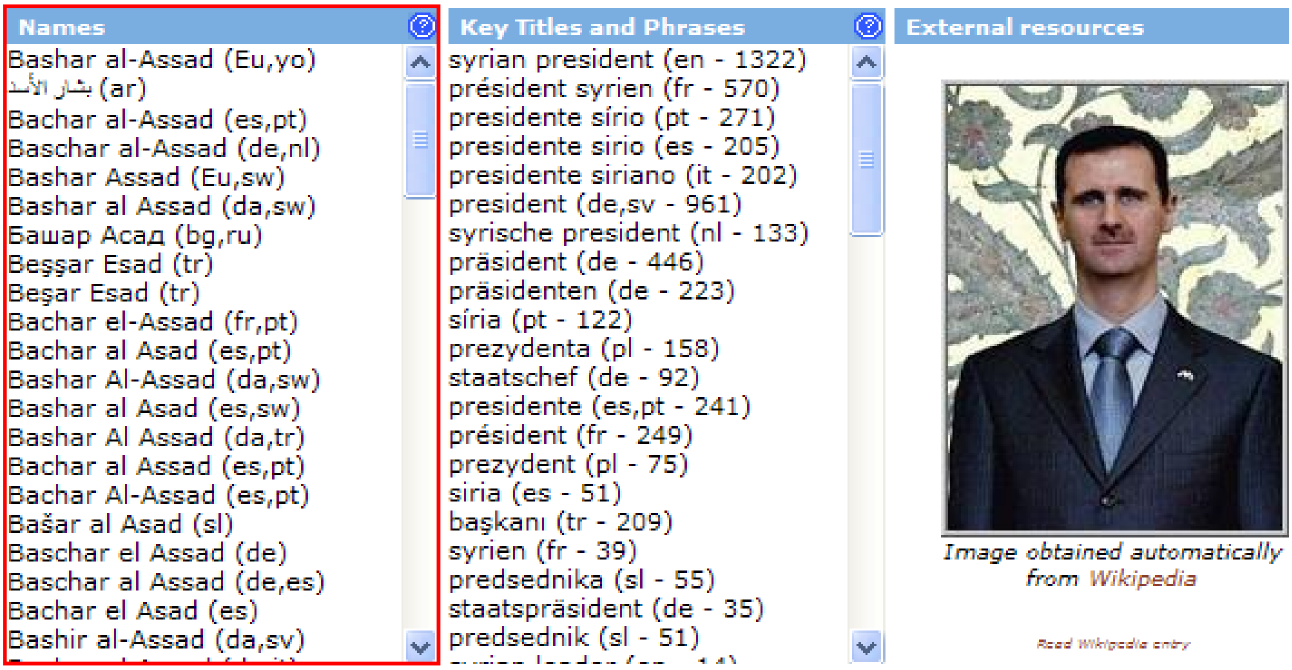




Bashar Assad

Bashar al-Assad; بشار الأسد; Bachar al-Assad; Baschar al-Assad; Bashar al Assad; Башар Асад; Beşşar Esad; Beşar Esad; Bachar el-Assad; Bachar al Asad; Bashar Al-Assad; Bashar al Asad; Bashar Al Assad; Bachar al Assad; Bachar Al-Assad; Bašar al Asad; Baschar el Assad; Baschar al Assad; Bachar el Asad; Bashir al-Assad; Bashar el Assad; Baschar Assad; Başar Esad; Beşşar Esed; Bashar al-Asad; Bachar Al Assad; Bashar Asad; Bachar Al Asad; Baszar el-Asad; Bachar El-Assad; Башар ал-Асад; Bachar Assad; Bachar El-Asad; Bachar el Assad; Basjar al-Assad; Bashar el-Assad; Bashar el Asad; Baschar Al-Assad; Bashar El Assad; Bashar Al-Asad; Baschir al-Assad; Basher Assad; Bachar al-Asad; Bachar El Assad; Bashar Al Asad; Bašar Asad; Baar al Asad; Bascher el Assad; Baschar el-Assad; Beşir Esad; Bashir al Assad; Başer Esad; Bashar al-Asssad; Bashir Al Assad; Bachar Asad; Basher al Assad; Beşar Essad; Baszira Asada; Baschir Assad; Beshar Assad; Bashar El-Asad; Basher al-Assad; Bashir el Assad; Bachar El Asad; Beşar el Esad; Bachar el-Asad; Baar Asad; Baschar al-Asad; Bashar El Asad; Bashar el-Asad; Baschar al-Asssad; Башар Accaд; Bashr al Assad; Bachir el Assad; Bashir el Asad; Bashir Al-Assad; Basjar Assad; Bashir al-Aassad;

Cross-lingual similarity — Entity names



Names

Bashar al-Assad (Eu,yo)
بشار الأسد (ar)
Bachar al-Assad (es,pt)
Baschar al-Assad (de,nl)
Bashar Assad (Eu,sw)
Bashar al Assad (da,sw)
Башар Асад (bg,ru)
Beşşar Esad (tr)
Beşar Esad (tr)
Bachar el-Assad (fr,pt)
Bachar al Asad (es,pt)
Bashar Al-Assad (da,sw)
Bashar al Asad (es,sw)
Bashar Al Assad (da,tr)
Bachar al Assad (es,pt)
Bachar Al-Assad (es,pt)
Bašar al Asad (sl)
Baschar el Assad (de)
Baschar al Assad (de,es)
Bachar el Asad (es)
Bashir al-Assad (da,sv)
Bashar al-Assad (da,it)

Key Titles and Phrases

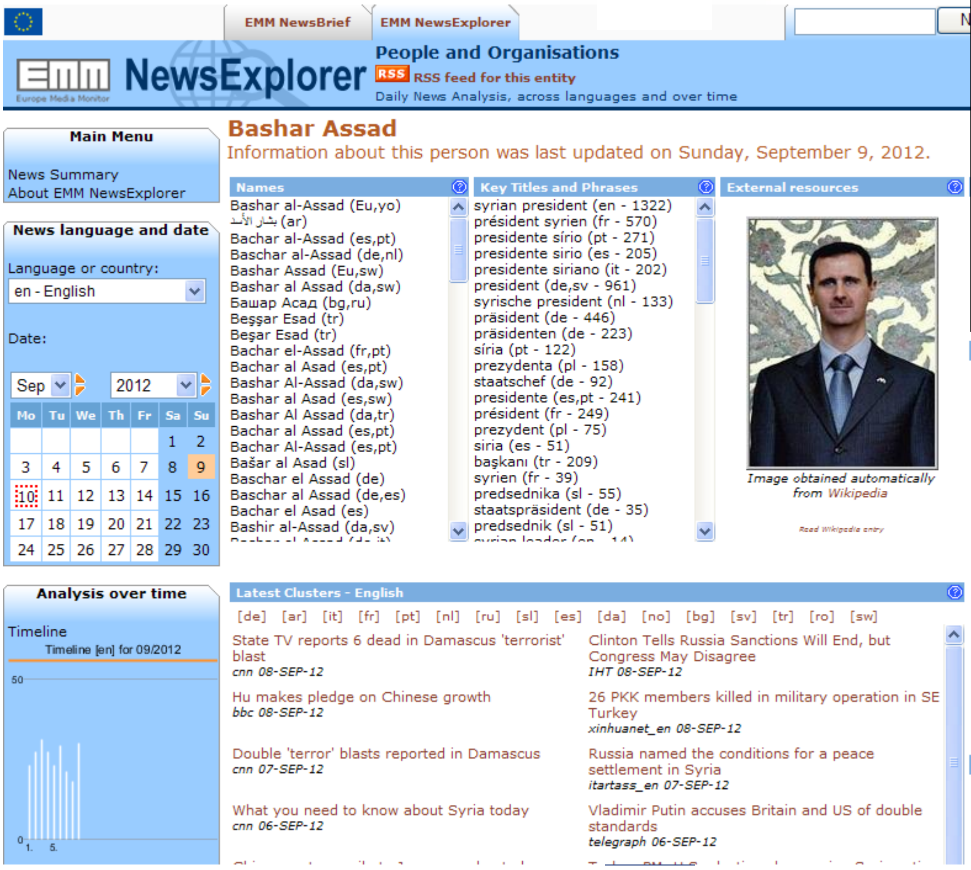
syrian president (en - 1322)
président syrien (fr - 570)
presidente sírio (pt - 271)
presidente sirio (es - 205)
presidente siriano (it - 202)
president (de,sv - 961)
syrische president (nl - 133)
präsident (de - 446)
präsidenten (de - 223)
síria (pt - 122)
prezydenta (pl - 158)
staatschef (de - 92)
presidente (es,pt - 241)
président (fr - 249)
prezydent (pl - 75)
siria (es - 51)
başkanı (tr - 209)
syrien (fr - 39)
predsednika (sl - 55)
staatspräsident (de - 35)
predsednik (sl - 51)
syrian leader (en - 14)

External resources

Image obtained automatically from Wikipedia

Read Wikipedia entry

Cross-lingual similarity — Entity names (2)



EMM NewsBrief | EMM NewsExplorer

EMM NewsExplorer
Europe Media Monitor

People and Organisations
RSS RSS feed for this entity
Daily News Analysis, across languages and over time

Main Menu
News Summary
About EMM NewsExplorer

News language and date
Language or country: en - English
Date: Sep 2012

Mo	Tu	We	Th	Fr	Sa	Su
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

Analysis over time
Timeline
Timeline [en] for 09/2012

Bashar Assad

Information about this person was last updated on Sunday, September 9, 2012.

Names

Bashar al-Assad (Eu,yo)
بشار الأسد (ar)
Bachar al-Assad (es,pt)
Baschar al-Assad (de,nl)
Bashar Assad (Eu,sw)
Bashar al Assad (da,sw)
Башар Асад (bg,ru)
Beşşar Esad (tr)
Beşar Esad (tr)
Bachar el-Assad (fr,pt)
Bachar al Asad (es,pt)
Bashar Al-Assad (da,sw)
Bashar al Asad (es,sw)
Bashar Al Assad (da,tr)
Bachar al Assad (es,pt)
Bachar Al-Assad (es,pt)
Bašar al Asad (sl)
Baschar el Assad (de)
Baschar al Assad (de,es)
Bachar el Asad (es)
Bashir al-Assad (da,sv)

Key Titles and Phrases

syrian president (en - 1322)
président syrien (fr - 570)
presidente sírio (pt - 271)
presidente sirio (es - 205)
presidente siriano (it - 202)
president (de,sv - 961)
syrische president (nl - 133)
präsident (de - 446)
präsidenten (de - 223)
síria (pt - 122)
prezydenta (pl - 158)
staatschef (de - 92)
presidente (es,pt - 241)
président (fr - 249)
prezydent (pl - 75)
siria (es - 51)
başkanı (tr - 209)
syrien (fr - 39)
predsednika (sl - 55)
staatspräsident (de - 35)
predsednik (sl - 51)

External resources

Image obtained automatically from Wikipedia

Read Wikipedia entry

Latest Clusters - English

[de] [ar] [it] [fr] [pt] [nl] [ru] [sl] [es] [da] [no] [bg] [sv] [tr] [ro] [sw]

State TV reports 6 dead in Damascus 'terrorist' blast
cnn 08-SEP-12

Hu makes pledge on Chinese growth
bbc 08-SEP-12

Double 'terror' blasts reported in Damascus
cnn 07-SEP-12

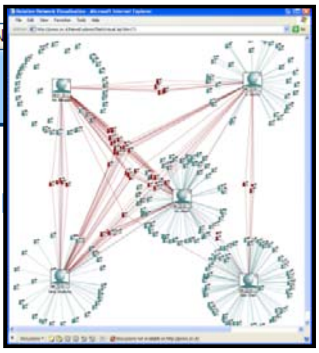
What you need to know about Syria today
cnn 06-SEP-12

Clinton Tells Russia Sanctions Will End, but Congress May Disagree
IHT 08-SEP-12

26 PKK members killed in military operation in SE Turkey
xinhuanet_en 08-SEP-12

Russia named the conditions for a peace settlement in Syria
itartass_en 07-SEP-12

Vladimir Putin accuses Britain and US of double standards
telegraph 06-SEP-12



Related People

Kofi Annan (2693)
Ban Ki Moon (1806)
Sergei Lavrov (1507)
Hillary Rodham Clinton (1474)
Barack Obama (1354)
Recep Tayyip Erdogan (981)
Nabil Elaraby (887)
Vladimir Putin (838)
Muammar Gaddafi (808)
Ahmet Davutoglu (805)
Alain Juppé (785)
Abdul Rahman (710)
Walid Muallem (645)
Nicolas Sarkozy (543)
William Hague (531)
Catherine Ashton (511)
Bourhan Ghalioun (509)
Susan Rice (445)
Guido Westerwelle (434)
David Cameron (413)
Victoria Nuland (402)
François Hollande (390)
Laurent Fabius (384)
Robert Mood (379)
Mahmoud Ahmadinejad (372)
Navi Pillay (370)

Associated People

Amer Alkhatib (2.0)
Mustafo al Dabija (1.2)
Vladimir Petrovich Kochyev (1.1)
Vladimir Kuyev (1.1)
Bourhan Ghalioun (1.1)

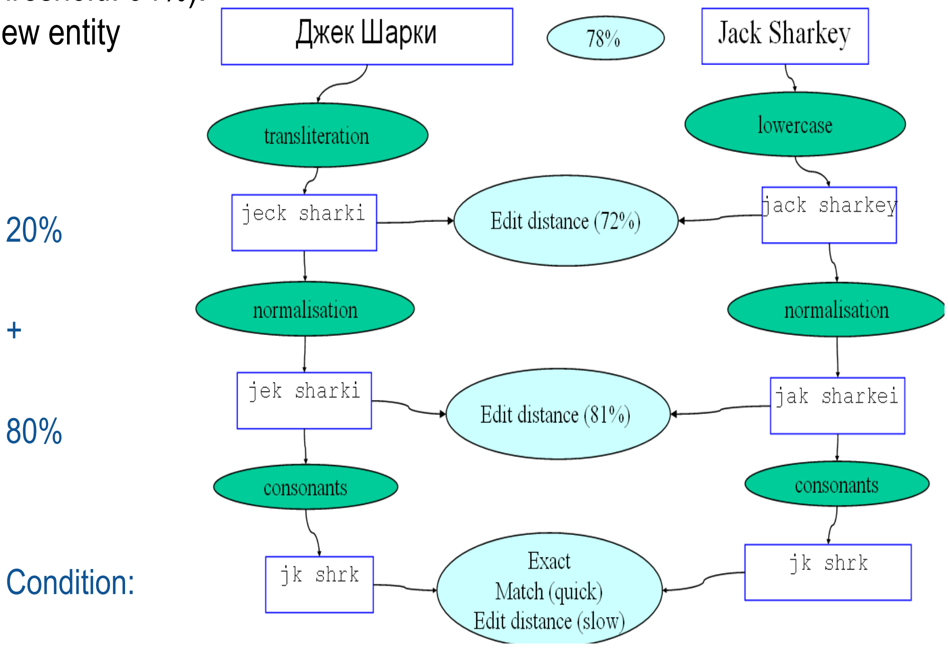
Multilingual NER

en	death of former Prime Minister Rafik Hariri, blamed by many opposition
es	asesinato del exprimer 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

Merging name variants

Steinberger Ralf & Bruno Pouliquen (2009). **Cross-lingual Named Entity Recognition.** In: Satoshi Sekine & Elisabete Ranchhod (eds.): Named Entities - Recognition, Classification and Use, Benjamins Current Topics, Volume 19, pp. 137-164. John Benjamins Publishing Company. ISBN 978-90-272-8922 3.

- For all newly found name forms, detect whether they are a variant of an existing NE:
  - Transliteration;
  - Normalisation, using ~30 hand-written rules and removing vowels;
  - Calculate similarity (threshold: 94%).
  - Below threshold → new entity



Джек Шарки

78%

Jack Sharkey

transliteration

lowercase

20%

jeck sharki

Edit distance (72%)

jack sharkey

normalisation

normalisation

+

jek sharki

Edit distance (81%)

jak sharkei

80%

consonants

consonants

Condition:

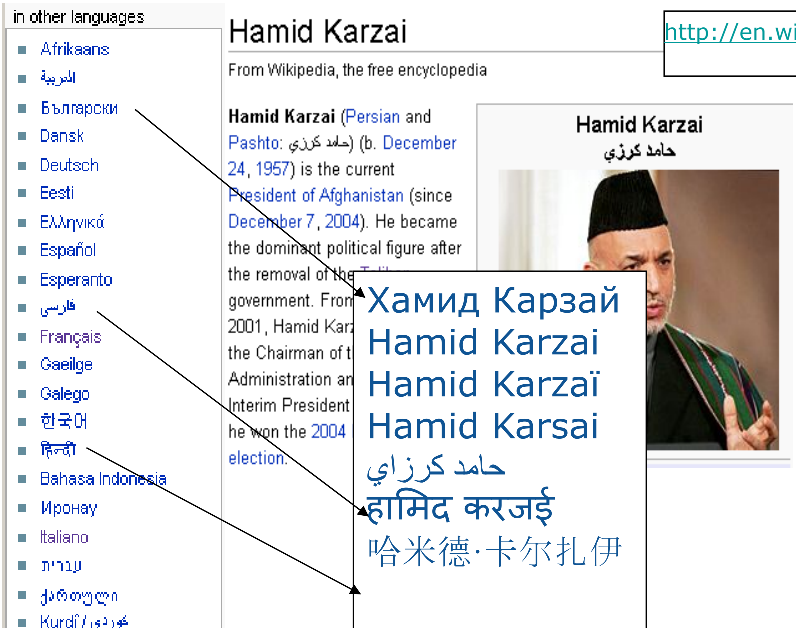
jk shrk

Exact
Match (quick)
Edit distance (slow)

jk shrk

Add Wikipedia variants

- For frequent or highly visible names, manually launch a Wikipedia mining process.
  - Check for each variant of a name whether there is a Wikipedia entry.
  - New name variants, *in all scripts*, will be recognised in new EMM articles.



in other languages

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

Hamid Karzai

From Wikipedia, the free encyclopedia

Hamid Karzai (Persian and Pashto: حامد کرزی) (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 Tali... government. From ... 2001, Hamid Karz... the Chairman of t... Administration an... Interim President ... he won the 2004 election.

Hamid Karzai
حامد کرزي

Хамид Карзай
Hamid Karzai
Hamid Karzaï
Hamid Karsai
حامد كرزاي
हामिद करजई
哈米德·卡尔扎伊

http://en.wikipedia.org/wiki/Hamid_Karzai

Freely available resource JRC-Names

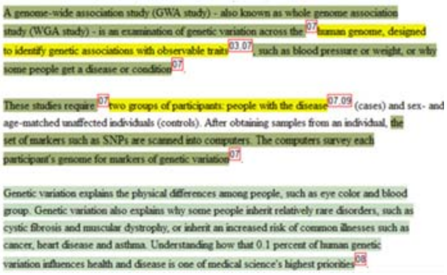
Name variant list, including across scripts, and software to recognise names in text

معمر القذافي; Mouammar Kadhafi; Muammar al-Gaddafi; Moammar Gadhafi; Muammar Gheddafi; Муамар Кадафи; Muammar Kadhafi; Muammar Kaddafi; Muammer Kaddafi; Muamar Gadafi; معمر قذافي; Moamerja Gadafija; Muammar Kadafi; Muammar el Gaddafi; Муамар Каддафи; Muamar el Gadafi; Moammar Gaddafi; Moamar Gaddafi; Moamer Kadhafi; Muammar Gadafi; Moamer Gadafi; Mouammar Khadafi; Moammar Kadhafi; Muammar Gadaffi; Muammar Khadaffi; Muammar Khaddafi; Muammar Qaddafi; Muhammar Gheddafi; Muammar al Gaddafi; Moammar Gadaffi; Muamar Kadafi; Муаммар Каддафи; Moamer Gathafi; Muammar Khadafi; Mouammar Kaddafi; Muamar Kadhafi; Muamar al Gadafi; Muammar el-Qaddafi; Muammar Gadafy; Muammar Kadaffi; Muammar Gadhafi; Moamer Gaddafi; Muammar al-Ghadhafi; Muamar Gaddafi; Muammar Ghaddafi; Muamar Khadafi; Muammar Ghadhafi; Muammar al-Gadafi; Muammar al-Qadhafi; Mouammar El Kadhafi; Muammar Qadhafi; Muammer Gadaffi; Moammar Gheddafi; Mouamar Kadhafi; Mouamar Khadafi; Moamer Kadaffi; Moammar al-Qadhafi; Moamer Qadhafi; Moamar Kadhafi; Moammar Khadafi; Moamar Gadafi; Moammar Qaddafi; Muammer Gaddafi; Muammar el-Gaddafi; Moeammar Kadhafi; Mummar Gaddafi; Muammar al-Qathafi; Muammar al-Kadhafi; Muammar Al-Kaddafi; Muammar Al-Qadhafi; Moammar Khadaffi; Muammar al-Qaddafi; Mouammar Al Kadhafi; Moammar Ghadafi; Muammar Al Gaddafi; Moammar Kaddafi; Moammar al-Kadhafi; Mouammar El-Kadhafi; Moammar Khaddafi; Moammar Qadhafi; Muammar al-Gathafi; Muammar Ghadaffi; Muhammar Gaddafi; Muammar Gaddaffi; Muammar el Gadafi; Muammar Abu Minyar al-Gaddafi; Muammar al-Kadafi; Muhamar Kadafi; Mouamar Kaddafi; Moammer Gaddafi; Muammar Al-Gaddafi; Muammar al-Khadafi; Mouammar El Khaddafi; Muammar Gadhaffi; Моамар Кадафи; Muamar Al Gadafi; Mouammar

Steinberger Ralf, Bruno Pouliquen, Mijail Kabadjov & Erik van der Goot (2011). **JRC-Names: A freely available, highly multilingual named entity resource.** Proceedings of the 8th International Conference Recent Advances in Natural Language Processing (**RANLP'2011**), pp. 104-110. Hissar, Bulgaria, 12-14 September 2011. (**PDF**)

Agenda

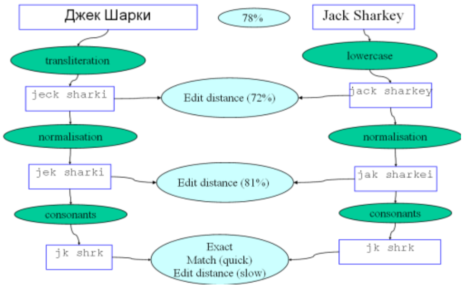
- EC-Joint Research Centre (JRC) – Who we are
- Monolingual plagiarism detection (PD) work at the JRC
- Cross-lingual similarity calculation at the JRC
  - Named entity (NE) matching across languages
  - Linking related news items across languages
  - Identifying translations of documents
- JRC's multilingual tools and resources
- Summary

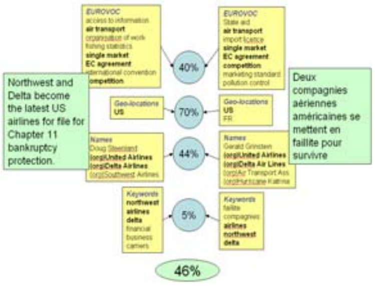


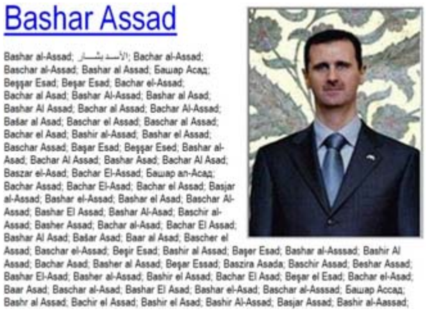
A genome-wide association study (GWA study) - also known as whole genome association study (WGA study) - is an examination of genetic variation across the 07 human genome, designed to identify genetic associations with observable traits 03,07, such as blood pressure or weight, or why some people get a disease or condition 07.

These studies require 07 two groups of participants: people with the disease 07,09 (cases) and sex- and age-matched unaffected individuals (controls). After obtaining samples from an individual, the set of markers such as SNPs are scanned into computers. The computers survey each participant's genome for markers of genetic variation 07.

Genetic variation explains the physical differences among people, such as eye color and blood group. Genetic variation also explains why some people inherit relatively rare disorders, such as cystic fibrosis and muscular dystrophy, or inherit an increased risk of common illnesses such as cancer, heart disease and asthma. Understanding how that 0.1 percent of human genetic variation influences health and disease is one of medical science's highest priorities 08.





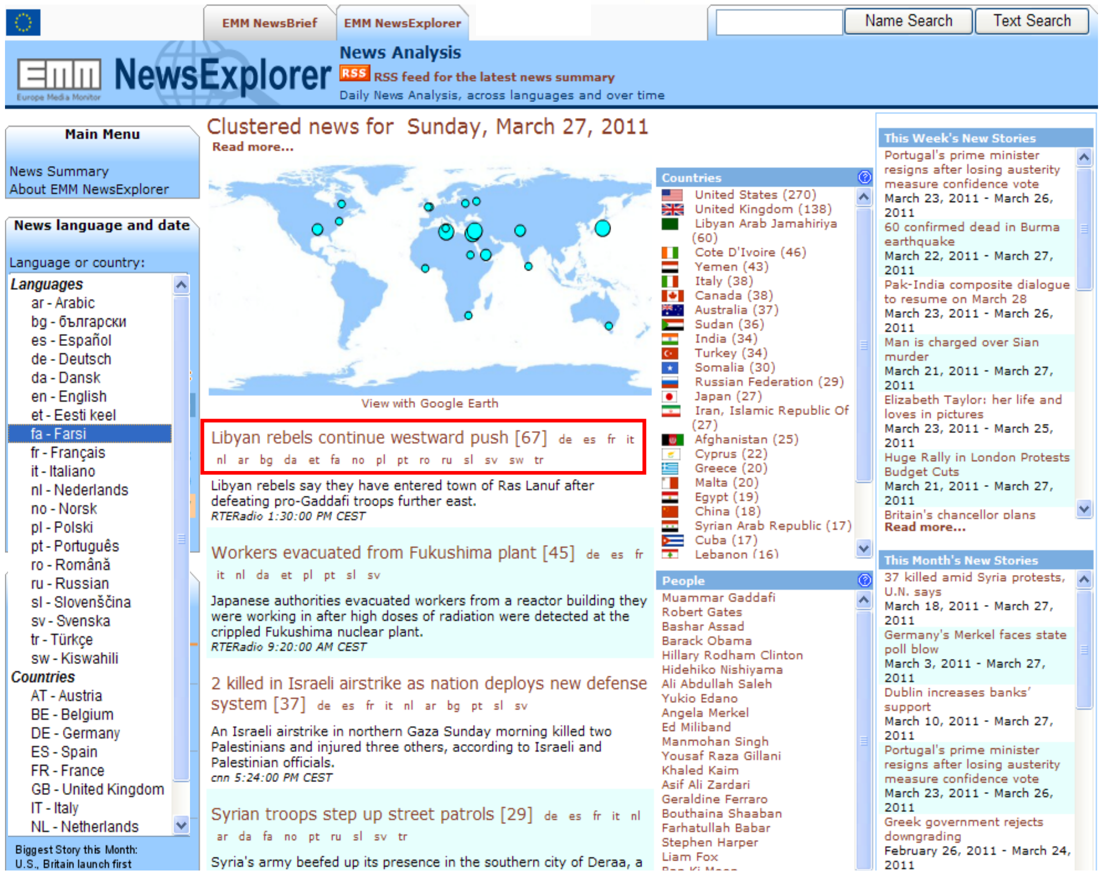


Bashar Assad

Bashar al-Assad; الأسد بشار; Bachar al-Assad; Baschar al-Assad; Bashar al Assad; Башар Асад; Beşşar Esad; Beşar Esad; Bachar el-Assad; Bachar al Asad; Bashar Al-Assad; Bashar al Asad; Bashar Al Assad; Bachar al Assad; Bachar Al-Assad; Bašar al Asad; Baschar el Assad; Baschar al Assad; Bachar el Asad; Bashir al-Assad; Bashar el Assad; Baschar Assad; Başar Esad; Beşşar Esed; Bashar al-Asad; Bachar Al Assad; Bashar Asad; Bachar Al Asad; Baszar el-Asad; Bachar El-Assad; Башар ал-Асад; Bachar Assad; Bachar El-Asad; Bachar el Assad; Basjar al-Assad; Bashar el-Assad; Bashar el Asad; Baschar Al-Assad; Bashar El Assad; Bashar Al-Asad; Baschir al-Assad; Basher Assad; Bachar al-Asad; Bachar El Assad; Bashar Al Asad; Bašar Asad; Baar al Asad; Bascher el Assad; Baschar el-Assad; Beşir Esad; Bashir al Assad; Başer Esad; Bashar al-Asssad; Bashir Al Assad; Bachar Asad; Basher al Assad; Beşar Essad; Baszira Asada; Baschir Assad; Beshar Assad; Bashar El-Asad; Basher al-Assad; Bashir el Assad; Bachar El Asad; Beşar el Esad; Bachar el-Asad; Baar Asad; Baschar al-Asad; Bashar El Asad; Bashar el-Asad; Baschar al-Asssad; Башар Ассад; Bashr al Assad; Bachir el Assad; Bashir el Asad; Bashir Al-Assad; Basjar Assad; Bashir al-Aassad;

Cross-lingual similarity

Documents



News Analysis

RSS RSS feed for the latest news summary

Daily News Analysis, across languages and over time

Main Menu

News Summary
About EMM NewsExplorer

News language and date

Language or country:

Languages
ar - Arabic
bg - български
es - Español
de - Deutsch
da - Dansk
en - English
et - Eesti keel
fa - Farsi
fr - Français
it - Italiano
nl - Nederlands
no - Norsk
pl - Polski
pt - Português
ro - Română
ru - Russian
sl - Slovenščina
sv - Svenska
tr - Türkçe
sw - Kiswahili
Countries
AT - Austria
BE - Belgium
DE - Germany
ES - Spain
FR - France
GB - United Kingdom
IT - Italy
NL - Netherlands

Biggest Story this Month:
U.S., Britain launch first

Clustered news for Sunday, March 27, 2011

Read more...

View with Google Earth

Libyan rebels continue westward push [67] de es fr it nl ar bg da et fa no pl pt ro ru sl sv sw tr

Libyan rebels say they have entered town of Ras Lanuf after defeating pro-Gaddafi troops further east.
RTERadio 1:30:00 PM CEST

Workers evacuated from Fukushima plant [45] de es fr it nl da et pl pt sl sv

Japanese authorities evacuated workers from a reactor building they were working in after high doses of radiation were detected at the crippled Fukushima nuclear plant.
RTERadio 9:20:00 AM CEST

2 killed in Israeli airstrike as nation deploys new defense system [37] de es fr it nl ar bg pt sl sv

An Israeli airstrike in northern Gaza Sunday morning killed two Palestinians and injured three others, according to Israeli and Palestinian officials.
cnn 5:24:00 PM CEST

Syrian troops step up street patrols [29] de es fr it nl ar da fa no pt ru sl sv tr

Syria's army beefed up its presence in the southern city of Deraa, a

Countries

United States (270)
United Kingdom (138)
Libyan Arab Jamahiriya (60)
Cote D'Ivoire (46)
Yemen (43)
Italy (38)
Canada (38)
Australia (37)
Sudan (36)
India (34)
Turkey (34)
Somalia (30)
Russian Federation (29)
Japan (27)
Iran, Islamic Republic Of (27)
Afghanistan (25)
Cyprus (22)
Greece (20)
Malta (20)
Egypt (19)
China (18)
Syrian Arab Republic (17)
Cuba (17)
Lebanon (16)

People

Muammar Gaddafi
Robert Gates
Bashar Assad
Barack Obama
Hillary Rodham Clinton
Hidehiko Nishiyama
Ali Abdullah Saleh
Yukio Edano
Angela Merkel
Ed Miliband
Manmohan Singh
Yousaf Raza Gillani
Khaled Kaim
Asif Ali Zardari
Geraldine Ferraro
Bouthaina Shaaban
Farhatullah Babar
Stephen Harper
Liam Fox

This Week's New Stories

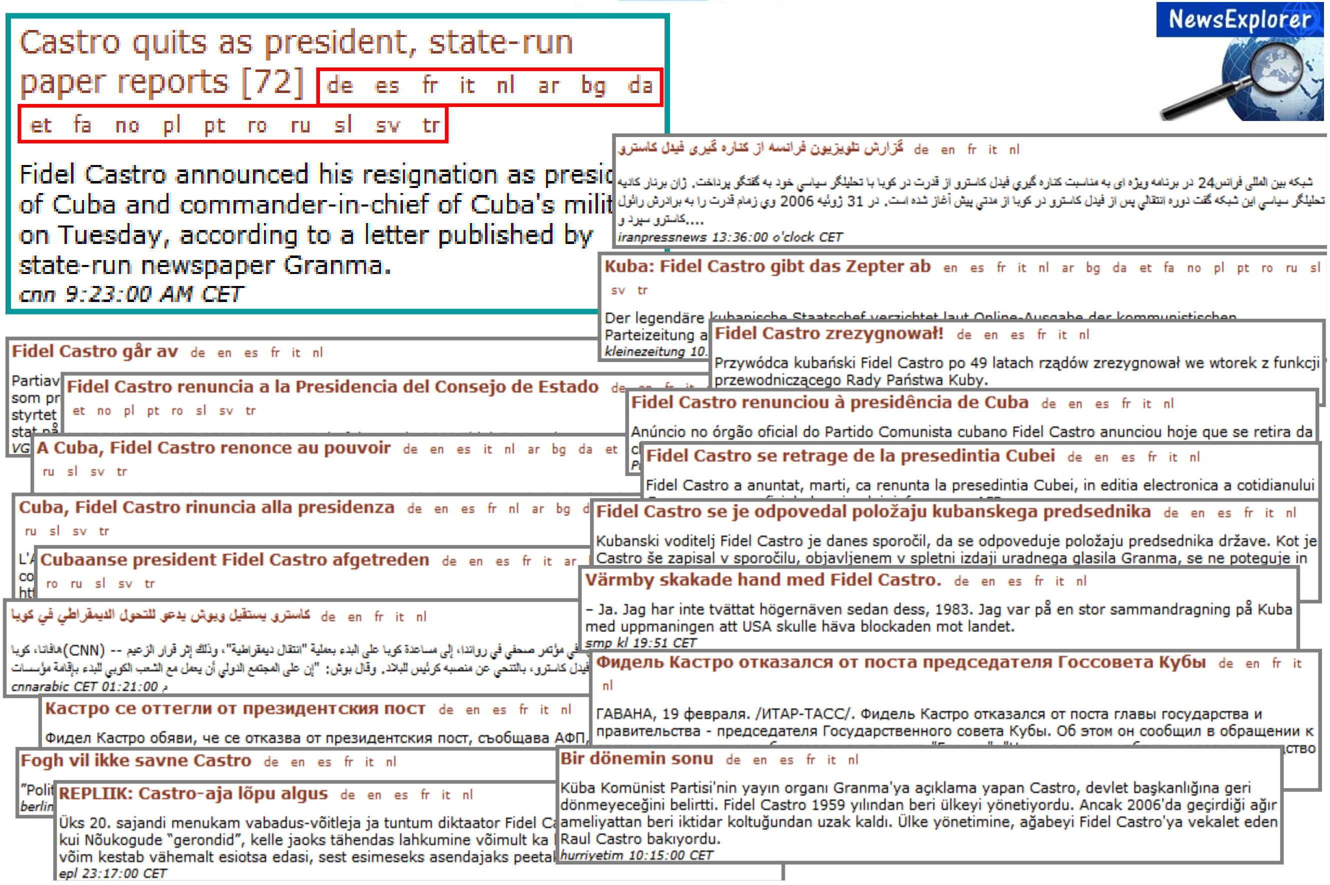
Portugal's prime minister resigns after losing austerity measure confidence vote
March 23, 2011 - March 26, 2011
60 confirmed dead in Burma earthquake
March 22, 2011 - March 27, 2011
Pak-India composite dialogue to resume on March 28
March 23, 2011 - March 26, 2011
Man is charged over Sian murder
March 21, 2011 - March 27, 2011
Elizabeth Taylor: her life and loves in pictures
March 23, 2011 - March 25, 2011
Huge Rally in London Protests Budget Cuts
March 21, 2011 - March 27, 2011
Britain's chancellor plans
Read more...

This Month's New Stories

37 killed amid Syria protests, U.N. says
March 18, 2011 - March 27, 2011
Germany's Merkel faces state poll blow
March 3, 2011 - March 27, 2011
Dublin increases banks' support
March 10, 2011 - March 27, 2011
Portugal's prime minister resigns after losing austerity measure confidence vote
March 23, 2011 - March 26, 2011
Greek government rejects downgrading
February 26, 2011 - March 24, 2011

live

Cross-lingual similarity   Documents (2)



Castro quits as president, state-run paper reports [72] de es fr it nl ar bg da et fa no pl pt ro ru sl sv tr

Fidel Castro announced his resignation as presid
of Cuba and commander-in-chief of Cuba's milit
on Tuesday, according to a letter published by
state-run newspaper Granma.
cnn 9:23:00 AM CET

گزارش تلویزیون فرانسه از کناره گیری فیدل کاسترو de en fr it nl

شبکه بین المللی فرانس24 در برنامه ویژه ای به مناسبت کناره گیری فیدل کاسترو از قدرت در کوبا با تحلیلگر سیاسی خود به گفتگو پرداخت. ژان برنار کادیه
تحلیلگر سیاسی این شبکه گفت دوره انتقالی پس از فیدل کاسترو در کوبا از مدتی پیش آغاز شده است. در 31 ژوئیه 2006 وی زمام قدرت را به برادرش راؤول
کاسترو سپرد و....
iranpressnews 13:36:00 o'clock CET

Kuba: Fidel Castro gibt das Zepter ab en es fr it nl ar bg da et fa no pl pt ro ru sl sv tr

Der legendäre kubanische Staatschef verzichtet laut Online-Ausgabe der kommunistischen
Parteizeitung a
kleinezeitung 10.

Fidel Castro zrezygnował! de en es fr it nl

Przywódca kubański Fidel Castro po 49 latach rządów zrezygnował we wtorek z funkcji
przewodniczącego Rady Państwa Kuby.

Fidel Castro går av de en es fr it nl

Partiav
som pr
styrtet
stat på
VG

Fidel Castro renuncia a la Presidencia del Consejo de Estado de en fr it
et no pl pt ro sl sv tr

Fidel Castro renunciou à presidência de Cuba de en es fr it nl

Anúncio no órgão oficial do Partido Comunista cubano Fidel Castro anunciou hoje que se retira da
c
P

A Cuba, Fidel Castro renonce au pouvoir de en es it nl ar bg da et
ru sl sv tr

Fidel Castro se retrage de la presedintia Cubei de en es fr it nl

Fidel Castro a anuntat, marti, ca renunta la presedintia Cubei, in editia electronica a cotidianului

Cuba, Fidel Castro rinuncia alla presidenza de en es fr nl ar bg d
ru sl sv tr

L'A
co
htt

Fidel Castro se je odpovedal položaju kubanskega predsednika de en es fr it nl

Kubanski voditelj Fidel Castro je danes sporočil, da se odpoveduje položaju predsednika države. Kot je
Castro še zapisal v sporočilu, objavljenem v spletni izdaji uradnega glasila Granma, se ne poteguje in

Cubaanse president Fidel Castro afgetreden de en es fr it ar
ro ru sl sv tr

Värmby skakade hand med Fidel Castro. de en es fr it nl

– Ja. Jag har inte tvättat högernäven sedan dess, 1983. Jag var på en stor sammandragning på Kuba
med uppmaningen att USA skulle häva blockaden mot landet.
smp kl 19:51 CET

كاسترو يستقيل وبوش يدعو للتحول الديمقراطي في كوبا de en fr it nl

في مؤتمر صحفي في رواندا، إلى مساعدة كوبا على البدء بعملية "انتقال ديمقراطية"، وذلك إثر قرار الزعيم -- (CNN)هافانا، كوبا
فيدل كاسترو، بالتنحي عن منصبه كرئيس للبلاد. وقال بوش: "إن على المجتمع الدولي أن يعمل مع الشعب الكوبي للبدء بإقامة مؤسسات
cnnarabic CET 01:21:00 م

Фидель Кастро отказался от поста председателя Госсовета Кубы de en fr it nl

ГАВАНА, 19 февраля. /ИТАР-ТАСС/. Фидель Кастро отказался от поста главы государства и
правительства - председателя Государственного совета Кубы. Об этом он сообщил в обращении к
ство

Кастро се оттегли от президентския пост de en es fr it nl

Фидел Кастро обяви, че се отказва от президентския пост, съобщава АФП,

Fogh vil ikke savne Castro de en es fr it nl

"Polit
berlin

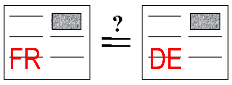
Bir dönemin sonu de en es fr it nl

Küba Komünist Partisi'nin yayın organı Granma'ya açıklama yapan Castro, devlet başkanlığına geri
dönmeyeceğini belirtti. Fidel Castro 1959 yılından beri ülkeyi yönetiyordu. Ancak 2006'da geçirdiği ağır
ameliyattan beri iktidar koltuğundan uzak kaldı. Ülke yönetimine, ağabeyi Fidel Castro'ya vekalet eden
Raul Castro bakıyordu.
hurriyetim 10:15:00 CET

REPLIIK: Castro-aja lõpu algus de en es fr it nl

Üks 20. sajandi menukam vabadus-võitleja ja tuntum diktaator Fidel Ca
kui Nõukogude "gerondid", kelle jaoks tähendas lahkumine võimult ka
võim kestab vähemalt esiotsa edasi, sest esimeseks asendajaks peetal
epl 23:17:00 CET

Cross-lingual Doc. Sim.   Introduction



- How to find out whether two texts in different languages are related?
- Most common approach: use MT or bilingual dictionaries to translate into English, then use monolingual methods to calculate similarity.
  - Using MT (e.g. Leek et al. 1999, Pinto et al. 2009);
  - Using bilingual dictionaries (e.g. Wactlar 1999, Urizar & Loinaz 2007)
- Automatically produce bilingual word associations for bilingual document representation and document similarity calculation, e.g.
  - Bilingual Lexical Semantic Analysis (LSA) (Landauer & Littman 1991)
  - Kernel Canonical Correlation Analysis (KCCA) (Vinokourov et al. 2002)
- Place documents in reference to position in comparable text collections (e.g. Wikipedia)
  - Cross-lingual Explicit Semantic Analysis (CL-ESA) (Potthast et al. 2008)

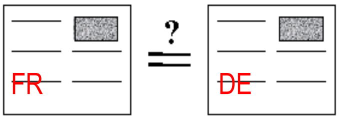
+ Achieved results are relatively good

- Bilingual approach is restricted to a few languages

Language pairs = N * (N -1) / 2 (N = number of languages)

20 NewsExplorer languages → 190 language pairs (380 language pair directions)!

Cross-lingual Doc. Sim. Our approach



- Alternative: use language-independent anchors:
  - Names of persons and organisations
  - Names of locations
  - Units of measurements:
    - Time
    - Speed
    - Temperature
    - Acceleration
  - Multilingual specialist dictionaries (Eurovoc for public administration, MeSH for medicine, etc.)
  - …
- Normalise these expressions

→Use as kind of an interlingua; no language pair-specific resource needed

- Similarly: Gupta et al. (2012) use Eurovoc and named entities

Steinberger Ralf, Bruno Pouliquen & Camelia Ignat (2004). **Providing cross-lingual information access with knowledge-poor methods.** In: Andrej Brodnik, Matjaž Gams & Ian Munro (eds.): **Informatica**. An international Journal of Computing and Informatics. Vol. 28-4, pp. 415-423. Special Issue 'Information Society in 2004'. ISSN: 0350-5596. The Slovene Society Informatika, Ljubljana, Slovenia.

Steinberger Ralf (2011). **A survey of methods to ease the development of highly multilingual Text Mining applications**. **Language Resources and Evaluation Journal**, Springer (DOI 10.1007/s10579-011-9165-9).

CL Document Similarity Language-independent anchors

Libyan rebels continue westward push [67] de es fr it
nl ar bg da et fa no pl pt ro ru sl sv sw tr

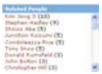
Language-independent features for multilingual document representation
No MT or bilingual dictionaries
20 languages

CLDS = α·S1 + β·S2 + γ·S3 + δ·S4

Sim1 (40%):
Multilingual Eurovoc subject domains

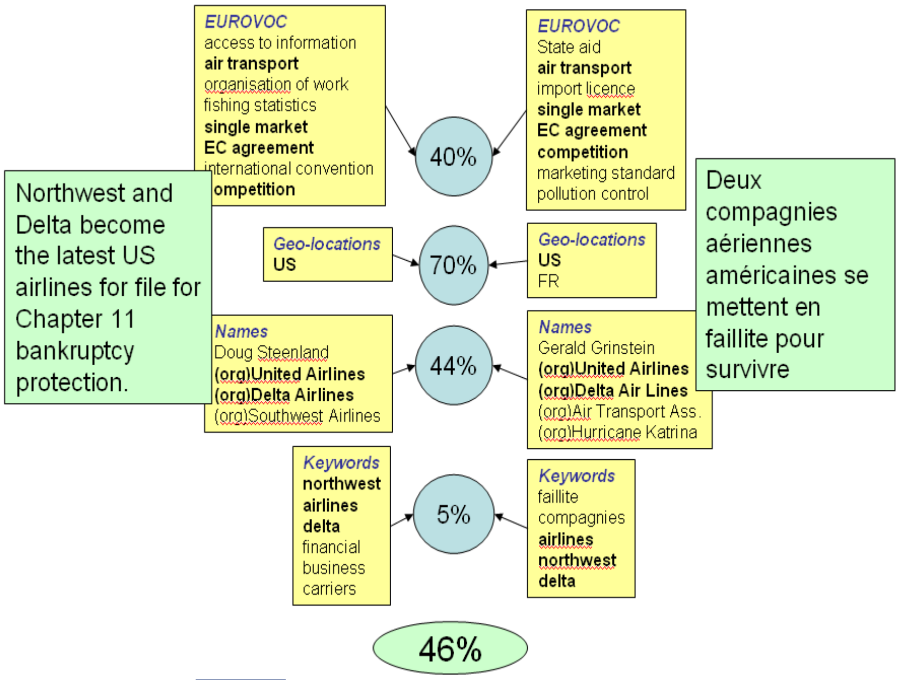
10.4184	*us*
1.5610	*gb*
1.5610	*il*
1.5610	*br*

Sim2 (30%):
Geo-locations



Sim3 (20%):
Names + variants

Sim4 (10%):
Cognates and numbers (without country score)



Northwest and Delta become the latest US airlines for file for Chapter 11 bankruptcy protection.

EUROVOC
access to information
air transport
organisation of work
fishing statistics
single market
EC agreement
international convention
ompetition

EUROVOC
State aid
air transport
import licence
single market
EC agreement
competition
marketing standard
pollution control

40%

Geo-locations
US

Geo-locations
US
FR

70%

Names
Doug Steenland
(org)United Airlines
(org)Delta Airlines
(org)Southwest Airlines

Names
Gerald Grinstein
(org)United Airlines
(org)Delta Air Lines
(org)Air Transport Ass.
(org)Hurricane Katrina

44%

Keywords
northwest
airlines
delta
financial
business
carriers

Keywords
faillite
compagnies
airlines
northwest
delta

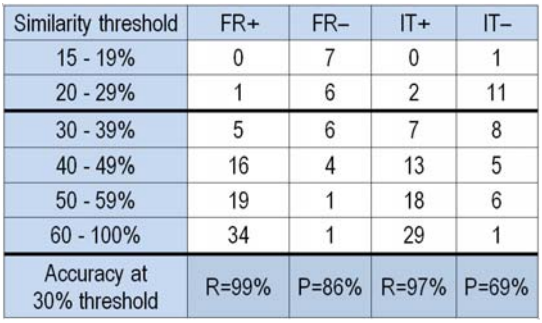
5%

Deux compagnies aériennes américaines se mettent en faillite pour survivre

46%

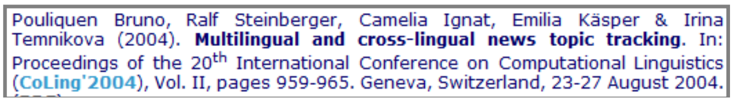
CL document similarity — Evaluation

- Task: evaluate manually the automatically proposed cross-lingual (CL) links
  - At various similarity threshold levels
  - ~25% of EN clusters had no cl links in FR and IT;
  - Only highest-scoring link was evaluated;
  - 30% threshold was finally chosen to ensure good Recall.



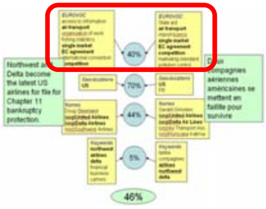
Similarity threshold	FR+	FR–	IT+	IT–
15 - 19%	0	7	0	1
20 - 29%	1	6	2	11
30 - 39%	5	6	7	8
40 - 49%	16	4	13	5
50 - 59%	19	1	18	6
60 - 100%	34	1	29	1
Accuracy at 30% threshold	R=99%	P=86%	R=97%	P=69%

+	Cluster was related
–	Cluster was not (so) related



Pouliquen Bruno, Ralf Steinberger, Camelia Ignat, Emilia Käsper & Irina Temnikova (2004). **Multilingual and cross-lingual news topic tracking**. In: Proceedings of the 20th International Conference on Computational Linguistics (**CoLing'2004**), Vol. II, pages 959-965. Geneva, Switzerland, 23-27 August 2004.

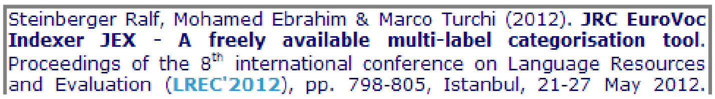
JRC EuroVoc Indexer JEX — ML EuroVoc Indexing



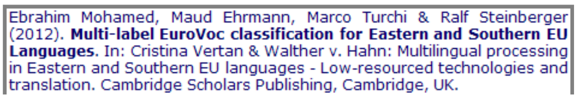
- JEX is multilingual multi-label classification software
  - Using the controlled vocabulary from EuroVoc (>6,000 classes)
- EuroVoc (http://eurovoc.europa.eu/)
  - is used for *manual* indexing by parliamentary libraries in EU institutions and in many EU countries
  - Exists in 22 official EU languages plus Basque, Catalan, Croatian, Russian and Serbian

EuroVoc Multilingual Thesaurus of the European Union

- JEX is freely downloadable from http://langtech.jrc.ec.europa.eu/Eurovoc.html;
  - Readily trained for 22 languages
- JEX includes software to re-train the system
  - Training data is included in the release;
  - Allows you to run your own experiments and compare results / improve.
  - You can train on your own data, using other thesauri.



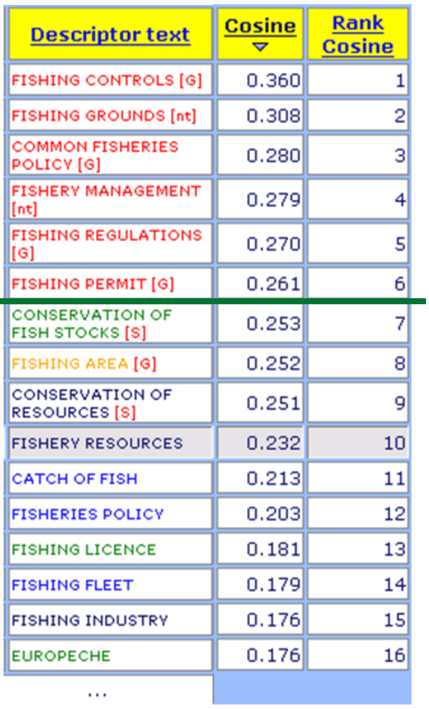
Steinberger Ralf, Mohamed Ebrahim & Marco Turchi (2012). **JRC EuroVoc Indexer JEX - A freely available multi-label categorisation tool**. Proceedings of the 8th international conference on Language Resources and Evaluation (**LREC'2012**), pp. 798-805, Istanbul, 21-27 May 2012.



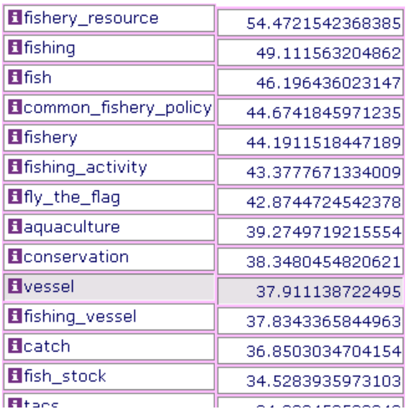
Ebrahim Mohamed, Maud Ehrmann, Marco Turchi & Ralf Steinberger (2012). **Multi-label EuroVoc classification for Eastern and Southern EU Languages**. In: Cristina Vertan & Walther v. Hahn: Multilingual processing in Eastern and Southern EU languages - Low-resourced technologies and translation. Cambridge Scholars Publishing, Cambridge, UK.

JRC EuroVoc Indexer JEX (2)

- Method: Profile-based category-ranking
  - E.g. Result for a document with the title:
    Legislative resolution embodying Parliament's opinion on the proposal for a Council Regulation amending Regulation No 2847/93 establishing a control system applicable to the common fisheries policy
  - E.g. profile for the EuroVoc category FISHERY MANAGEMENT

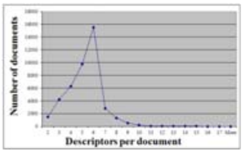


Descriptor text	Cosine	Rank Cosine
FISHING CONTROLS [G]	0.360	1
FISHING GROUNDS [nt]	0.308	2
COMMON FISHERIES POLICY [G]	0.280	3
FISHERY MANAGEMENT [nt]	0.279	4
FISHING REGULATIONS [G]	0.270	5
FISHING PERMIT [G]	0.261	6
CONSERVATION OF FISH STOCKS [S]	0.253	7
FISHING AREA [G]	0.252	8
CONSERVATION OF RESOURCES [S]	0.251	9
FISHERY RESOURCES	0.232	10
CATCH OF FISH	0.213	11
FISHERIES POLICY	0.203	12
FISHING LICENCE	0.181	13
FISHING FLEET	0.179	14
FISHING INDUSTRY	0.176	15
EUROPECHE	0.176	16
...		



fishery_resource	54.4721542368385
fishing	49.111563204862
fish	46.196436023147
common_fishery_policy	44.6741845971235
fishery	44.1911518447189
fishing_activity	43.3777671334009
fly_the_flag	42.8744724542378
aquaculture	39.2749719215554
conservation	38.3480454820621
vessel	37.911138722495
fishing_vessel	37.8343365844963
catch	36.8503034704154
fish_stock	34.5283935973103
tacs	

Evaluation:
P, R, F1 at *rank* 6.

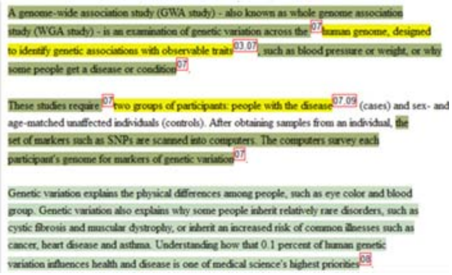


JEX evaluation for 22 languages

Language	Precision	Recall	F1	F1 dynamic rank	Categories / collection	Average categories trained	Stop words used (+MW)	Total number of documents	Document length (words) ± std-dev	All categories trained
BG	0.4619	0.5120	0.4857	0.4940	3780	2049.9	332	22696	786.96±2784.72	2147
CS	0.4689	0.5205	0.4933	0.4990	3691	2035.7	137	22830	890.66±3317.10	2129
DA	0.4747	0.5491	0.5092	0.5170	4226	2655	858	41727	561.87±1875.19	2752
DE	0.4732	0.5485	0.5081	0.5187	4230	**2683.9**	1793	41676	566.67±1945.44	**2783**
EL	0.4632	0.5369	0.4973	0.5118	4214	2486.4	105	41103	778.05±2379.45	2605
EN	0.4801	0.5547	0.5147	0.5227	4229	2324.1	**1972 (+545)**	41672	309.28±1176.39	2430
ES	0.4801	0.5545	0.5147	0.5188	4221	2297	481 (+264)	41397	547.63±1819.14	2406
ET	0.4828	0.5358	0.5079	0.5139	3662	2047.8	1533	21989	652.22±2193.32	2147
FI	0.4654	0.5341	0.4974	0.5081	4103	2528.8	92	38293	756.70±2565.81	2634
FR	0.4776	0.5536	0.5128	0.5227	**4234**	2588.7	1180	**41989**	663.33±2204.28	2688
HU	**0.5121**	**0.5654**	**0.5374**	**0.5444**	3585	**1688.5**	1228 (+709)	20838	551.66±1977.14	**1788**
IT	0.4713	0.5464	0.5061	0.5151	**4234**	2584.4	219	41838	764.57±2808.31	2688
LT	0.4920	0.5454	0.5174	0.5239	3635	1945.7	1199	21505	644.53±2724.18	2046
LV	0.4659	0.5175	0.4904	0.4968	3690	2011	14	22803	894.59±3012.39	2106
MT	**0.4200**	**0.4545**	**0.4366**	**0.4416**	**3584**	1762.3	**6**	**17858**	1016.99±2685.11	1864
NL	0.4803	0.5562	0.5155	0.5257	4232	2610.2	1414	41816	581.94±1819.66	2713
PL	0.4794	0.5311	0.5039	0.5077	3648	1967.1	125	22004	841.81±2795.35	2066
PT	0.4756	0.5493	0.5098	0.5237	4209	2560.6	1152	41142	700.46±2138.01	2663
RO	0.4550	0.5043	0.4784	0.4817	3887	2109.3	1504 (+48)	25023	994.17±3083.16	2206
SK	0.4705	0.5204	0.4942	0.4995	3645	1938.4	364	21406	872.53±3241.50	2050
SL	0.4840	0.5341	0.5078	0.5205	3685	2013.1	2068	22289	627.56±2669.38	2119
SV	0.4787	0.5473	0.5107	0.5194	4109	2546.4	1093	38198	609.82±2365.26	2649

Agenda

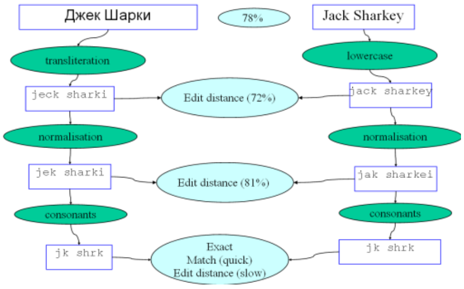
- EC-Joint Research Centre (JRC) – Who we are
- Monolingual plagiarism detection (PD) work at the JRC
- Cross-lingual similarity calculation at the JRC
  - Named entity (NE) matching across languages
  - Linking related news items across languages
  - Identifying translations of documents
- JRC's multilingual tools and resources
- Summary

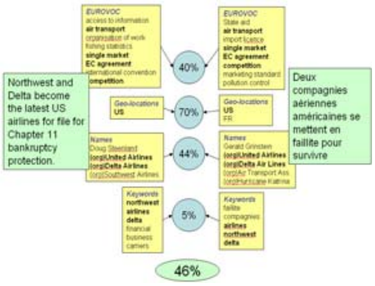


A genome-wide association study (GWA study) - also known as whole genome association study (WGA study) - is an examination of genetic variation across the [07] human genome, designed to identify genetic associations with observable traits [03.07], such as blood pressure or weight, or why some people get a disease or condition [07].

These studies require [07] two groups of participants: people with the disease [07.09] (cases) and sex- and age-matched unaffected individuals (controls). After obtaining samples from an individual, the set of markers such as SNPs are scanned into computers. The computers survey each participant's genome for markers of genetic variation [07].

Genetic variation explains the physical differences among people, such as eye color and blood group. Genetic variation also explains why some people inherit relatively rare disorders, such as cystic fibrosis and muscular dystrophy, or inherit an increased risk of common illnesses such as cancer, heart disease and asthma. Understanding how that 0.1 percent of human genetic variation influences health and disease is one of medical science's highest priorities [08].





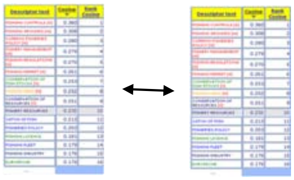
Bashar Assad

Bashar al-Assad; الأسد بشار; Bachar al-Assad; Baschar al-Assad; Bashar al Assad; Башар Асад; Beşşar Esad; Beşar Esad; Bachar el-Assad; Bachar al Asad; Bashar Al-Assad; Bashar al Asad; Bashar Al Assad; Bachar al Assad; Bachar Al-Assad; Bašar al Asad; Baschar el Assad; Baschar al Assad; Bachar el Asad; Bashir al-Assad; Bashar el Assad; Baschar Assad; Başar Esad; Beşşar Esed; Bashar al-Asad; Bachar Al Assad; Bashar Asad; Bachar Al Asad; Baszar el-Asad; Bachar El-Assad; Башар ал-Асад; Bachar Assad; Bachar El-Asad; Bachar el Assad; Basjar al-Assad; Bashar el-Assad; Bashar el Asad; Baschar Al-Assad; Bashar El Assad; Bashar Al-Asad; Baschir al-Assad; Basher Assad; Bachar al-Asad; Bachar El Assad; Bashar Al Asad; Bašar Asad; Baar al Asad; Bascher el Assad; Baschar el-Assad; Beşir Esad; Bashir al Assad; Başer Esad; Bashar al-Asssad; Bashir Al Assad; Bachar Asad; Basher al Assad; Beşar Essad; Baszira Asada; Baschir Assad; Beshar Assad; Bashar El-Asad; Basher al-Assad; Bashir el Assad; Bachar El Asad; Beşar el Esad; Bachar el-Asad; Baar Asad; Baschar al-Asad; Bashar El Asad; Bashar el-Asad; Baschar al-Asssad; Башар Ассад; Bashr al Assad; Bachir el Assad; Bashir el Asad; Bashir Al-Assad; Basjar Assad; Bashir al-Aassad;

Translation spotting using EuroVoc indexing

Task: find Spanish translations of English source document in a parallel text collection

by calculating the cosine similarity between document's EuroVoc vectors.



Is the document's translation the most similar document in the other language?

Precision at rank 1.

	Search space	Without length factor	With length factor
Simple document similarity (DS)	820 Es	90.61%	96.83%
DS considering the length of documents	820 Es	00.12%	01.71%
Different text type	795 Es	84.28%	90.31%
Mixed-language search space	410 Es + 410 En	69.68%	81.91%
DS correcting mono-lingual bias (83%)	410 Es + 410 En	92.91%	96.82%

Pouliquen Bruno, Ralf Steinberger & Camelia Ignat (2003). **Automatic Identification of Document Translations in Large Multilingual Document Collections.** In: Proceedings of the International Conference *Recent Advances in Natural Language Processing* (RANLP'2003). Borovets, Bulgaria,

Translation spotting

Is there a translation?

- Setting a threshold, juggling Precision and Recall

Test bed	Average similarity	Threshold	Recall	Noise (1-Precision)
Set T1 (820)	0.82	0.70	90%	2.2%
Set T2 (795)	0.79	0.70	76.5%	5%

- Searching for a translation where there is none:
  - Searching in T2 for documents of T1

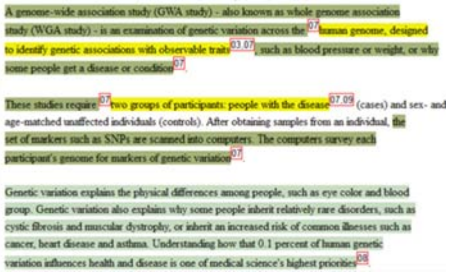
  → 4.15% noise

- Best threshold depends on:
  - Document set
  - Requirement: high recall or high precision?

Pouliquen Bruno, Ralf Steinberger & Camelia Ignat (2003). **Automatic Identification of Document Translations in Large Multilingual Document Collections.** In: Proceedings of the International Conference *Recent Advances in Natural Language Processing* (RANLP'2003). Borovets, Bulgaria,

Agenda

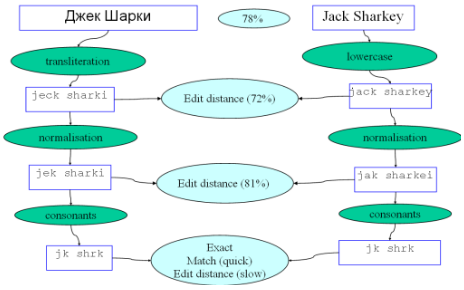
- EC-Joint Research Centre (JRC) – Who we are
- Monolingual plagiarism detection (PD) work at the JRC
- Cross-lingual similarity calculation at the JRC
  - Named entity (NE) matching across languages
  - Linking related news items across languages
  - Identifying translations of documents
- JRC's multilingual tools and resources
- Summary

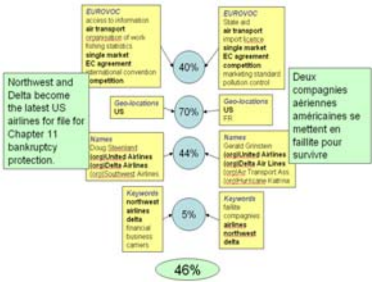


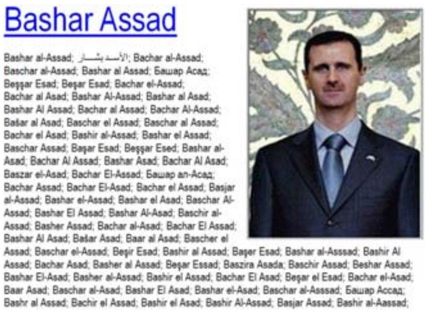
A genome-wide association study (GWA study) - also known as whole genome association study (WGA study) - is an examination of genetic variation across the 07 human genome, designed to identify genetic associations with observable traits 03.07, such as blood pressure or weight, or why some people get a disease or condition 07.

These studies require 07 two groups of participants: people with the disease 07.09 (cases) and sex- and age-matched unaffected individuals (controls). After obtaining samples from an individual, the set of markers such as SNPs are scanned into computers. The computers survey each participant's genome for markers of genetic variation 07.

Genetic variation explains the physical differences among people, such as eye color and blood group. Genetic variation also explains why some people inherit relatively rare disorders, such as cystic fibrosis and muscular dystrophy, or inherit an increased risk of common illnesses such as cancer, heart disease and asthma. Understanding how that 0.1 percent of human genetic variation influences health and disease is one of medical science's highest priorities 08.







Bashar Assad

Bashar al-Assad; الأسد بشار; Bachar al-Assad; Baschar al-Assad; Bashar al Assad; Башар Асад; Beşşar Esad; Beşar Esad; Bachar el-Assad; Bachar al Asad; Bashar Al-Assad; Bashar al Asad; Bashar Al Assad; Bachar al Assad; Bachar Al-Assad; Bašar al Asad; Baschar el Assad; Baschar al Assad; Bachar el Asad; Bashir al-Assad; Bashar el Assad; Baschar Assad; Başar Esad; Beşşar Esed; Bashar al-Asad; Bachar Al Assad; Bashar Asad; Bachar Al Asad; Baszar el-Asad; Bachar El-Assad; Башар ал-Асад; Bachar Assad; Bachar El-Asad; Bachar el Assad; Basjar al-Assad; Bashar el-Assad; Bashar el Asad; Baschar Al-Assad; Bashar El Assad; Bashar Al-Asad; Baschir al-Assad; Basher Assad; Bachar al-Asad; Bachar El Assad; Bashar Al Asad; Bašar Asad; Baar al Asad; Bascher el Assad; Baschar el-Assad; Beşir Esad; Bashir al Assad; Başer Esad; Bashar al-Asssad; Bashir Al Assad; Bachar Asad; Basher al Assad; Beşar Essad; Baszira Asada; Baschir Assad; Beshar Assad; Bashar El-Asad; Basher al-Assad; Bashir el Assad; Bachar El Asad; Beşar el Esad; Bachar el-Asad; Baar Asad; Baschar al-Asad; Bashar El Asad; Bashar el-Asad; Baschar al-Asssad; Башар Ассад; Bashr al Assad; Bachir el Assad; Bashir el Asad; Bashir Al-Assad; Basjar Assad; Bashir al-Aassad;

Freely available resources

European Commission

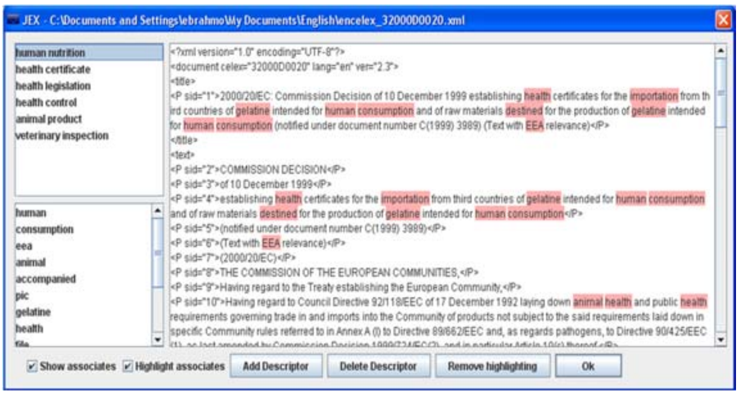
JRC Eurovoc Indexer JEX

EuroVoc Multilingual Thesaurus of the European Union

- Software to multi-label classify documents according to the multilingual Eurovoc thesaurus:

- 22 languages, thousands of categories;
- JEX uses machine learning;
- Tool can be re-trained on users' documents, also for non-Eurovoc categories
- User interface and command-line options
- Tool is used for cross-lingual linking of news in EMM-NewsExplorer
- In use by the Spanish *Congress of Deputies* for interactive indexing since 2006.

Congreso de los Diputados



JEX - C:\Documents and Settings\ebrahmo\My Documents\English\encelex_32000D0020.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<document celex="32000D0020" lang="en" ver="2.3">
<title>
<P sid="1">2000/20/EC: Commission Decision of 10 December 1999 establishing health certificates for the importation from th
ird countries of gelatine intended for human consumption and of raw materials destined for the production of gelatine intended
for human consumption (notified under document number C(1999) 3989) (Text with EEA relevance)</P>
</title>
<text>
<P sid="2">COMMISSION DECISION</P>
<P sid="3">of 10 December 1999</P>
<P sid="4">establishing health certificates for the importation from third countries of gelatine intended for human consumption
and of raw materials destined for the production of gelatine intended for human consumption</P>
<P sid="5">(notified under document number C(1999) 3989)</P>
<P sid="6">(Text with EEA relevance)</P>
<P sid="7">(2000/20/EC)</P>
<P sid="8">THE COMMISSION OF THE EUROPEAN COMMUNITIES,</P>
<P sid="9">Having regard to the Treaty establishing the European Community,</P>
<P sid="10">Having regard to Council Directive 92/118/EEC of 17 December 1992 laying down animal health and public health
requirements governing trade in and imports into the Community of products not subject to the said requirements laid down in
specific Community rules referred to in Annex A (I) to Directive 89/662/EEC and, as regards pathogens, to Directive 90/425/EEC
```

Freely available resources

European Commission

JRC-Names

Name variant list and software to recognise names

- JRC-Names: a highly multilingual named entity resource
  (names and their many spelling variants, including across scripts):

  - Collected by analysing up to 150,000 news articles per day in up to 20 languages since 2004
  - Augmented with cross-script variants from Wikipedia, resulting in currently:
  - ~500K person and organisation names and their spelling variants
  - In 27 scripts and many more languages

معمر القذافي; Mouammar Kadhafi; Muammar al-Gaddafi; Moammar Gadhafi; Muammar Gheddafi; Муамар Кадафи; Muammar Kadhafi; Muammar Kaddafi; Muammer Kaddafi; Muamar Gadafi; معمر قذافي; Moamerja Gadafija; Muammar Kadafi; Muammar el Gaddafi; Муамар Каддафи; Muamar el Gadafi; Moammar Gaddafi; Moamar Gaddafi; Moamer Kadhafi; Muammar Gadafi; Moamer Gadafi; Mouammar Khadafi; Moammar Kadhafi; Muammar Gadaffi; Muammar Khadaffi; Muammar Khaddafi; Muammar Qaddafi; Muhammar Gheddafi; Muammar al Gaddafi; Moammar Gadaffi; Muamar Kadafi; Муаммар Каддафи; Moamer Gathafi; Muammar Khadafi; Mouammar Kaddafi; Muamar Kadhafi; Muamar al Gadafi; Muammar el-Qaddafi; Muammar Gadafy; Muammar Kadaffi; Muammar Gadhafi; Moamer Gaddafi; Muammar al-Ghadhafi; Muamar Gaddafi; Muammar Ghaddafi; Muamar Khadafi; Muammar Ghadhafi; Muammar al-Gadafi; Muammar al-Qadhafi; Mouammar El Kadhafi; Muammar Qadhafi; Muammer Gadaffi; Moammar Gheddafi; Mouamar Kadhafi; Mouamar Khadafi; Moamer Kadaffi; Moammar al-Qadhafi; Moamer Qadhafi; Moamar Kadhafi; Moammar Khadafi; Moamar Gadafi; Moammar Qaddafi; Muammer Gaddafi; Muammar el-Gaddafi; Moeammar Kadhafi; Mummar Gaddafi; Muammar al-Qathafi; Muammar al-Kadhafi; Muammar Al-Kaddafi; Muammar Al-Qadhafi; Moammar Khadaffi; Muammar al-Qaddafi; Mouammar Al Kadhafi; Moammar Ghadafi; Muammar Al Gaddafi; Moammar Kaddafi; Moammar al-Kadhafi; Mouammar El-Kadhafi; Moammar Khaddafi; Moammar Qadhafi; Muammar al-Gathafi; Muammar Ghadaffi; Muhammar Gaddafi; Muammar Gaddaffi; Muammar el Gadafi; Muammar Abu Minyar al-Gaddafi; Muammar al-Kadafi; Muhamar Kadafi; Mouamar Kaddafi; Moammer Gaddafi; Muammar Al-Gaddafi; Muammar al-Khadafi; Mouammar El Khaddafi; Muammar Gadhaffi; Моамар Кадафи; Muamar Al Gadafi; Mouammar

Freely available resources — Parallel corpora

European Commission

Possible uses:

- Train statistical machine translation software;
- Train multilingual vector space models (e.g. LSA or KCCA);
- Derive multilingual dictionaries;
- …

Data already available (22 languages each):

JRC-Acquis	full-text parallel corpus	(domain: mostly legal; agreements; contracts)
DGT-TM	Translation Memory	(domain: mostly legal)
JEX data	full-text parallel corpus	(domain: legislation)

Forthcoming (25, 25, 23 languages):

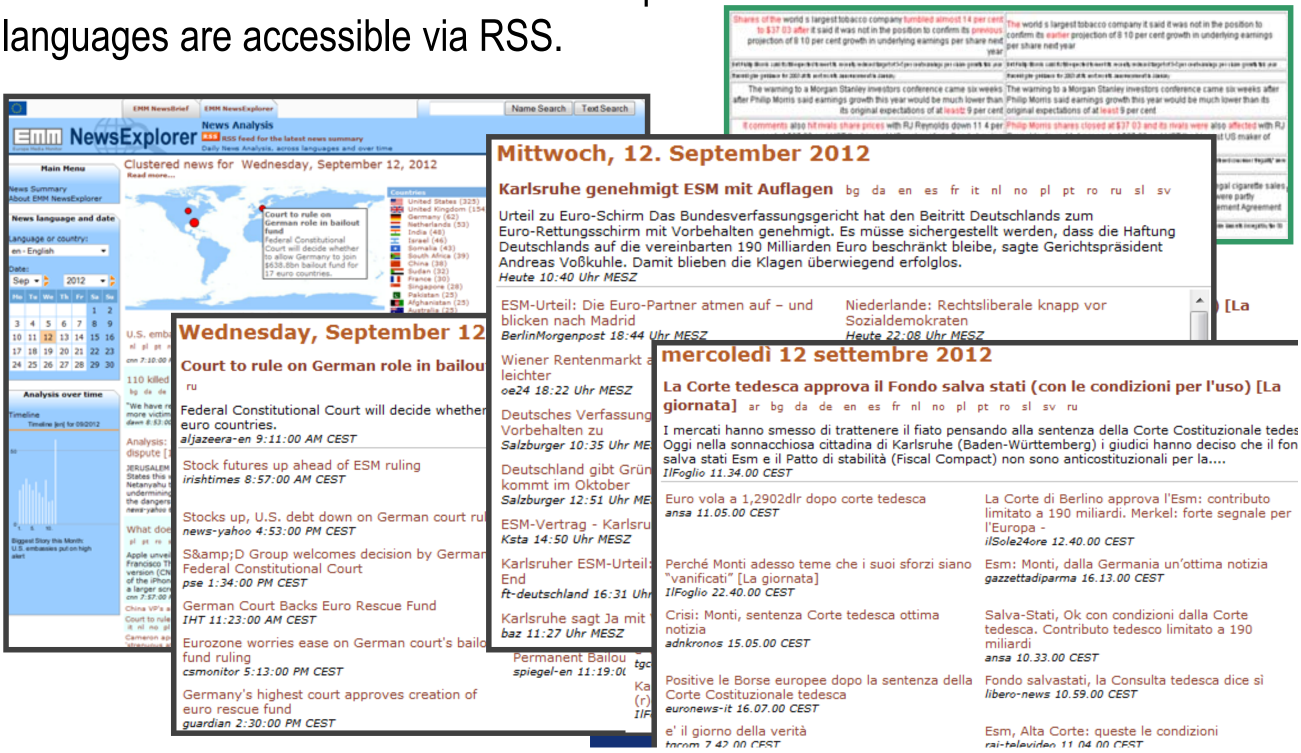
EAC-TM	Translation Memory	(domain: education and culture)
ECDC-TM	Translation Memory	(domain: public health and medicine)
DGT-Acquis	full-text parallel corpus	(domain: legal, administration and more)

•See: http://langtech.jrc.ec.europa.eu/JRC_Resources.html

Accessible resource — multilingual news clusters

NewsExplorer News Analysis RSS RSS feed for the latest news summary Daily News Analysis, across languages and over time

Meta-data for news clusters and their equivalences in other languages are accessible via RSS.



Mittwoch, 12. September 2012

Karlsruhe genehmigt ESM mit Auflagen bg da en es fr it nl no pl pt ro ru sl sv

Urteil zu Euro-Schirm Das Bundesverfassungsgericht hat den Beitritt Deutschlands zum Euro-Rettungsschirm mit Vorbehalten genehmigt. Es müsse sichergestellt werden, dass die Haftung Deutschlands auf die vereinbarten 190 Milliarden Euro beschränkt bleibe, sagte Gerichtspräsident Andreas Voßkuhle. Damit blieben die Klagen überwiegend erfolglos.
Heute 10:40 Uhr MESZ

ESM-Urteil: Die Euro-Partner atmen auf – und blicken nach Madrid
BerlinMorgenpost 18:44 Uhr MESZ

Niederlande: Rechtsliberale knapp vor Sozialdemokraten
Heute 22:08 Uhr MESZ

Wiener Rentenmarkt a leichter
oe24 18:22 Uhr MESZ

Deutsches Verfassung Vorbehalten zu
Salzburger 10:35 Uhr ME

Deutschland gibt Grün kommt im Oktober
Salzburger 12:51 Uhr ME

ESM-Vertrag - Karlsru
Ksta 14:50 Uhr MESZ

Karlsruher ESM-Urteil: End
ft-deutschland 16:31 Uhr

Karlsruhe sagt Ja mit
baz 11:27 Uhr MESZ

Permanent Bailou
spiegel-en 11:19:0

Wednesday, September 12

Court to rule on German role in bailou ru

Federal Constitutional Court will decide whether euro countries.
aljazeera-en 9:11:00 AM CEST

Stock futures up ahead of ESM ruling
irishtimes 8:57:00 AM CEST

Stocks up, U.S. debt down on German court rul
news-yahoo 4:53:00 PM CEST

S&amp;D Group welcomes decision by German Federal Constitutional Court
pse 1:34:00 PM CEST

German Court Backs Euro Rescue Fund
IHT 11:23:00 AM CEST

Eurozone worries ease on German court's bailo fund ruling
csmonitor 5:13:00 PM CEST

Germany's highest court approves creation of euro rescue fund
guardian 2:30:00 PM CEST

mercoledì 12 settembre 2012

La Corte tedesca approva il Fondo salva stati (con le condizioni per l'uso) [La giornata] ar bg da de en es fr nl no pl pt ro sl sv ru

I mercati hanno smesso di trattenere il fiato pensando alla sentenza della Corte Costituzionale tedes Oggi nella sonnacchiosa cittadina di Karlsruhe (Baden-Württemberg) i giudici hanno deciso che il fon salva stati Esm e il Patto di stabilità (Fiscal Compact) non sono anticostituzionali per la....
IlFoglio 11.34.00 CEST

Euro vola a 1,2902dlr dopo corte tedesca
ansa 11.05.00 CEST

La Corte di Berlino approva l'Esm: contributo limitato a 190 miliardi. Merkel: forte segnale per l'Europa -
ilSole24ore 12.40.00 CEST

Perché Monti adesso teme che i suoi sforzi siano "vanificati" [La giornata]
IlFoglio 22.40.00 CEST

Esm: Monti, dalla Germania un'ottima notizia
gazzettadiparma 16.13.00 CEST

Crisi: Monti, sentenza Corte tedesca ottima notizia
adnkronos 15.05.00 CEST

Salva-Stati, Ok con condizioni dalla Corte tedesca. Contributo tedesco limitato a 190 miliardi
ansa 10.33.00 CEST

Positive le Borse europee dopo la sentenza della Corte Costituzionale tedesca
euronews-it 16.07.00 CEST

Fondo salvastati, la Consulta tedesca dice sì
libero-news 10.59.00 CEST

e' il giorno della verità
tgcom 7.42.00 CEST

Esm, Alta Corte: queste le condizioni
rai-televideo 11.04.00 CEST

EMM NewsExplorer
News Analysis
RSS feed for the latest news summary
Name Search
Text Search
Main Menu
News Summary
About EMM NewsExplorer
News language and date
Language or country:
en - English
Date:
Sep
2012
Analysis over time
Clustered news for Wednesday, September 12, 2012
Court to rule on German role in bailout fund
Federal Constitutional Court will decide whether to allow Germany to join $638.8bn bailout fund for 17 euro countries.
United States (325)
United Kingdom (154)
Germany (62)
Netherlands (53)
India (48)
Israel (46)
Somalia (43)
South Africa (39)
China (38)
Sudan (32)
France (30)
Singapore (28)
Pakistan (25)
Afghanistan (25)
Australia (25)
The world s largest tobacco company it said it was not in the position to confirm its earlier projection of 8 10 per cent growth in underlying earnings per share next year
The warning to a Morgan Stanley investors conference came six weeks after Philip Morris said earnings growth this year would be much lower than its original expectations of at least 9 per cent
Philip Morris shares closed at $37 03 and its rivals were also affected

Summary

- Monolingual plagiarism detection work at the JRC
  - N-gram overlap; varying search and visualisation methods

- Cross-lingual similarity calculation at the JRC
  - Named entity matching across languages
  - Linking related news items across languages
  - Identifying translations of documents

- JRC's multilingual tools and resources
  - JRC-Names – multilingual name variant lists
  - JEX – EuroVoc subject domain classification
  - Parallel corpora: JRC-Acquis and various translation memories
  - Cross-lingual linking of news clusters in NewsExplorer