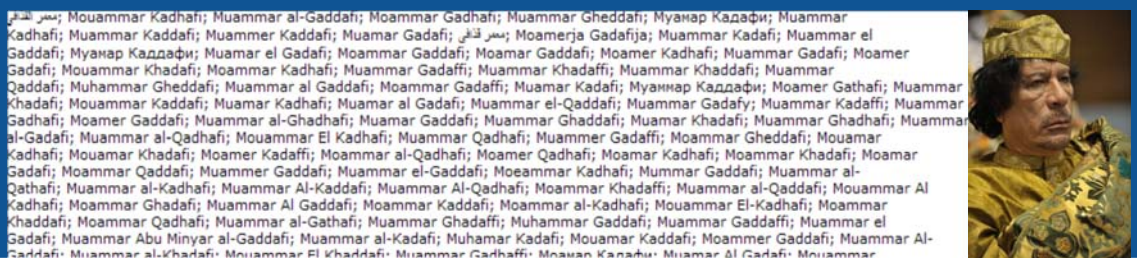
JRC-Names: A freely available multilingual name variant spelling dictionary

Ralf Steinberger
European Commission – Joint Research Centre (JRC)

Translation at the Frontiers of the Lexicon:
The New Fields of Terminology (TSIB'2013)

Brussels, 31 May 2013
Institut supérieur de traducteurs et interprètes, Haute École de Bruxelles

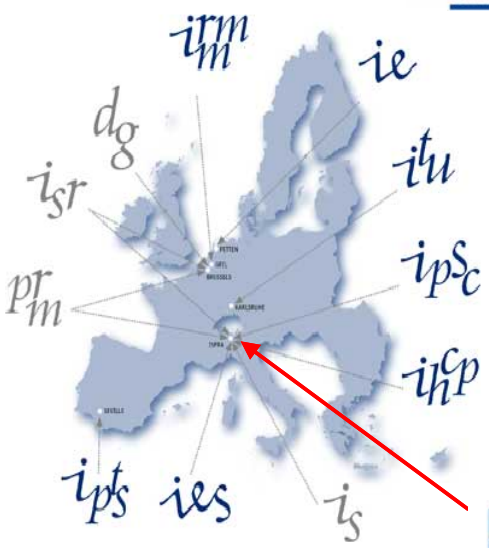


Agenda

- EC-JRC: Who we are and what we do
- What is JRC-Names; What can it be used for
- Related work: other named entity (NE) resources
- How JRC-Names was produced
- Statistics on JRC-Names
- Current work: multilingual acronym recognition
- Further multilingual linguistic resources
- Summary

Entity 100006
Giovanni+Bosco
John+Bosco
Don+Bosco
Jean+Bosco
Joao+Bosco
Sveti+Janez+Bosco
Jan+Bosco
Yohannes+Bosco
Šv+Jonas+Boskas
若望+鮑思高
Gioan+Bosco
Bosco+Szent+Janos
Iohannes+Bosco
若望鮑思高
Ioan+Bosco
Jan+Bosko
ヨハネボスコ
Don+Bosko
João+Bosco
Bosco+Szent+János
ヨハネ+ボスコ
Ivan+Bosco
若望·鮑思高
Juan+Bosco
Giovanni+Melchior+Bosco
นักบุญยอห์น+บอสโก
Johannes+Bosco
ヨハネ・ボスコ
Sant+Joan+Bosco
Иоанн+Боско
Sveti+Ivan+Bosco

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**)

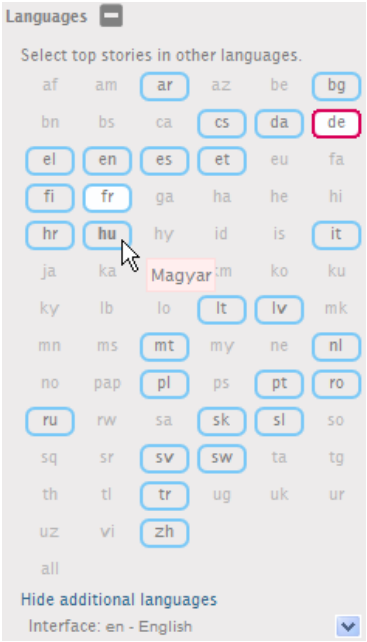
- European Commission
  (scientific-technical arm of public administration)
- Non-commercial
- Multi-disciplinary / multilingual

EMM news gathering — Facts and numbers

Europe Media Monitor (EMM) news gathering - A few facts

~ 175,000 articles / day in over 70 languages (>40 active)
~ 4000 Sources (world-wide, with focus on Europe)

- news sources (web portals)
- specialist medical sites
- ~ 20 commercial newswires
- 24/7, updated every 10 minutes



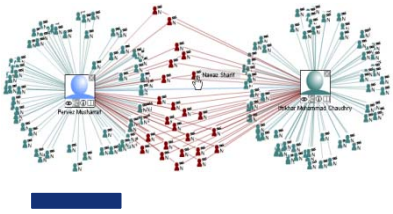
Articles are fed into the various EMM applications for analysis:

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.

EMM news analysis

Automatic analysis of the daily news:

- Group related news articles into clusters
- Categorise news into over 1000 subject domains
- Produce a variety of statistics
- Detect trends
- Early warning functionality per category and country.
- Link related news over time
- Link related news across languages
- Automatically extract information about *entities*:
  - Names of people, organisations, locations
  - Quotations by and about people („…“)
- Gather information about entities.
- Social networks of related people
- Event extraction
- Machine Translation
- Opinion Mining …



EMM users

EMM media monitoring users – wide coverage, world-wide

European Commission (most DGs) and other EU Institutions

EU Agencies:

- e.g. Public Health (ECDC), Food Safety (EFSA), Chemicals Bureau (ECHA), etc.

EU Member State organisations: e.g.

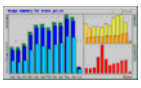
- Public Health,
- law enforcement authorities,
- parliaments,
- crisis management/humanitarian

International and extra-European organisations: e.g.

- various UN organisations
- Centres for Disease Prevention and Control in the US, Canada, China, …

The public:

- ~30,000 anonymous internet users of publicly accessible EMM systems.
- Combined between 1 and 2 Million hits per day



EMM mobile version

Coming soon. Free

Agenda

European Commission

- EC-JRC: Who we are and what we do
- What is JRC-Names; What can it be used for
- Related work: other named entity (NE) resources
- How JRC-Names was produced
- Statistics on JRC-Names
- Current work: multilingual acronym recognition
- Further multilingual linguistic resources
- Summary

Entity 100006

Giovanni+Bosco
John+Bosco
Don+Bosco
Jean+Bosco
Joao+Bosco
Sveti+Janez+Bosco
Jan+Bosco
Yohannes+Bosco
Sv+Jonas+Boskas
若望+鮑思高
Gioan+Bosco
Bosco+Szent+Janos
Iohannes+Bosco
若望鮑思高
Ioan+Bosco
Jan+Bosko
ヨハネボスコ
Don+Bosko
João+Bosco
Bosco+Szent+János
ヨハネ+ボスコ
Ivan+Bosco
若望·鮑思高
Juan+Bosco
Giovanni+Melchior+Bosco
นักบุญยอห์น+บอสโก
Johannes+Bosco
ヨハネ・ボスコ
Sant+Joan+Bosco
Иоанн+Боско
Sveti+Ivan+Bosco

What is JRC-Names

JRC-Names consists of:

- Lists of names and their many spelling variants,
  - ~205,000 person and organisation names plus
  - ~204,000 name spelling variants
  - In 27 scripts and many more languages (status Sept. 2011)
- Software to recognise these names in multilingual text, with offset and unique name identifier
- Download from http://ipsc.jrc.ec.europa.eu/?id=61

```
Entity 3202: UN
United+Nations
Nations+Unies
ONU
ಸಂಯುಕ್ತ+ರಾಷ್ಟ್ರ+ಸಂಸ್ಥೆ
Ujedinjeni narodi
FN
```

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

Possible uses of JRC-Names

Standardise name spellings in databases, text collections and the internet for improved retrieval (Stern & Sagot 2010)

Improve Machine Translation – names must be treated differently from other words (Babych & Hartley 2003; Steinberger & Pouliquen 2009)

Use as input to learn automatic transliteration rules (e.g. Pouliquen 2009)

Use output of JRC-Names as seeds to learn NER rules (e.g. Buchholz & van den Bosch 2000)

Social networks are less biased by national viewpoints if based on information extracted from multilingual texts

NER results are useful for other text mining tasks (opinion mining; co-reference resolution; summarisation; topic detection and tracking; cross-lingual linking of related documents across languages; …)

Variants found in 6h of EMM news!

id	Count	Names
262	1220	Muammar Gaddafi / Muammar al-Gaddafi / Mouammar Kadhafi / معمر القذافي / Muamar Gadafi / Muammer Kaddafi / Muammar Gheddafi / Муаммара Каддафи / Moammar Gadhafi / Муамара Каддафи / Muamar el Gadafi / Muammar Kadhafi / Muammar Kaddafi / Muammar Gaddafin / 卡扎菲 / Muammar al-Ghadhafi / Муамар Кадафи / معمر قذافي / Muammara Kaddáfi / Muammara Kadafiego / Muammar el Gaddafi / Muammar Kadafi / Moammar Kadhafi / Muamar al Gadafi / Muammar al Gaddafi / Moamerja Gadafija / Moamer Kadhafi / Муаммар Каддафи / Muammar Khadafi / Moammer Kadhafi / Moammar al-Kadhafi / Муамару Каддафи / Muammar Gadafy / Moammar Gaddafi / Muammar Qaddafi / Moamer Gadafi / Moammar Qaddafi / Moamer Gaddafi / Muammar Khaddafi / Муаммару Каддафи / Moamar Gadafi / Muammarui Gaddafi / Муамар Каддафи / Muamar Khadafi / Muammar Gadhafi / Muammar Khadaffi / لمعمر القذافي / Muammar al-Gadhafi / Muammar el Gadafi / Muamaro Kadhafi / معمر القذافي / Muammaro Gaddafi / Muammar Gadaffi / Muammar Kaddáfi / Mouammar El Gueddafi / Muammarui Gaddafiui / Muamar Kadhafi / Moamer Gadhafi / Muammaro Gaddafio / Muamar El Gadafi / Муамаром Каддафи / Muamaras Kadhafi / Muammar al-Gadafi / Муаммар Каддафі / Mouammar Kaddafi / Muamara Kaddafiego / Muamaras Gaddafi / Muammar Gadhaffi / Muammar Ghaddafi / Muhammar Gadaffi / Muamar Kadafi / Moammer Gadhafi / Muammar el-Qaddafi / 穆阿迈尔· 卡扎菲 / Moamerjem Gadafijem [b]
3202	987	United Nations / Naciones Unidas / Verenigde Naties / Nations unies / ONU / الأمم المتحدة / Onu / FN / Birleşmiş Milletler / Nações Unidas / 联合国 / Nazioni Unite / Организации Объединенных Наций / Nations Unies / Organizace spojených národů / سازمان ملل متحد / ONU. / ONU, / Organização das Nações Unidas / Nations-Unies / Onu, / Jungtinių Tautų Organizacija / FN. / Onu. / Liên Hiệp Quốc / Organización de las Naciones Unidas [b]
10101	701	EU / Unión Europea / European Union / Avrupa Birliği / União Europeia / EU. / Europese Unie / Union européenne / Ευρωπαϊκή Ένωση / EU / الاتحاد الأوروبي / EU, / Eiropas Savienība / Unii Europejskiej / Euroopa Liidu / Europäischen Union / Uniunea Europeană / اتحاديه اروپا / Európai Unió / Unión Europa / Unione Europea / Европейский Союз / Europos Sąjunga / Unione Europ / EÚ, / Europäische Union / UE / Euroopan unioni / Europska unija / Euroopa Liit / Unione europea / Unió Europea / EU? / EÚ / Unia Europejska / Європейський Союз / Union Européenne / Evropske Unije / Unjoni Ewropea / Europeiska unionen / EÚ. / האיחוד האירופי / UE, / union européenne [b]

Agenda

- EC-JRC: Who we are and what we do
- What is JRC-Names; What can it be used for
- Related work: other named entity (NE) resources
- How JRC-Names was produced
- Statistics on JRC-Names
- Current work: multilingual acronym recognition
- Further multilingual linguistic resources
- Summary

Entity 100006

Giovanni+Bosco
John+Bosco
Don+Bosco
...+...
Jean+Bosco
Joao+Bosco
Sveti+Janez+Bosco
Jan+Bosco
Yohannes+Bosco
Šv+Jonas+Boskas
若望+鮑思高
Gioan+Bosco
Bosco+Szent+Janos
Iohannes+Bosco
若望鮑思高
Ioan+Bosco
Jan+Bosko
ヨハネボスコ
Don+Bosko
João+Bosco
Bosco+Szent+János
ヨハネ+ボスコ
Ivan+Bosco
若望·鮑思高
Juan+Bosco
Giovanni+Melchior+Bosco
นักบุญยอห์น+บอสโก
Johannes+Bosco
ヨハネ・ボスコ
Sant+Joan+Bosco
Иоанн+Боско
Sveti+Ivan+Bosco

Other NE resources — Related work

Wentlant et al. (2008) – built a ml NE repository based on *Wikipedia* links and case information; 2.5 Mio English names, 250K German, 3K Swahili, …

Toral et al. (2008) – built Named Entity WordNet by searching NEs in *WordNet* and *Wikipedia*: 310K entities, including 278K persons

Stern & Sagot (2010) – exploit *French Wikipedia* and *GeoNames* to produce French resource: 263K person names + 883K variants.

Maurel (2009) –produced Prolexbase mostly manually: 75K entities of all types

→ Most resources are based on Wikipedia

- Strong at providing cross-lingual and cross-script variants;
- Offers only few other spelling variants:

JRC-Names contains mostly spelling variants from real-life text, enriched with Wikipedia – up to 413 variants for the same NE.

Agenda

- EC-JRC: Who we are and what we do
- What is JRC-Names; What can it be used for
- Related work: other named entity (NE) resources
- How JRC-Names was produced
- Statistics on JRC-Names
- Current work: multilingual acronym recognition
- Further multilingual linguistic resources
- Summary

Entity 100006
Giovanni+Bosco
John+Bosco
Don+Bosco
Jean+Bosco
Joao+Bosco
Sveti+Janez+Bosco
Jan+Bosco
Yohannes+Bosco
Św+Jonas+Boskas
若望+鮑思高
Gioan+Bosco
Bosco+Szent+Janos
Iohannes+Bosco
若望鮑思高
Ioan+Bosco
Jan+Bosko
ヨハネボスコ
Don+Bosko
João+Bosco
Bosco+Szent+János
ヨハネ+ボスコ
Ivan+Bosco
若望·鮑思高
Juan+Bosco
Giovanni+Melchior+Bosco
นักบุญยอห์น+บอสโก
Johannes+Bosco
ヨハネ・ボスコ
Sant+Joan+Bosco
Иоанн+Боско
Sveti+Ivan+Bosco

Multilingual NER in EMM — Brief overview

Lookup of most frequent ***known names*** and their variants in all languages

- Database contains over 1.2 million names + 225.000 variants (status July 2011)
- Including morphological (and other) variants by pre-generating inflection forms (Slovene example):
  Tony(a|o|u|om|em|m|ju|jem|ja)?\s+Blair(a|o|u|om|em|m|ju|jem|ja)

Guessing ***new names*** using empirically-derived *lexical patterns* in 20 languages.

- President, Minister, Head of State, Sir, American
- "death of", "[0-9]+-year-old", …
- Known first names + uppercase words

- Identification of a current average of 1,000 unknown names per day.
- Only names found repeatedly will become *known names* (error reduction).

Steinberger Ralf & Bruno Pouliquen (2007). **Cross-lingual Named Entity Recognition.** In: Satoshi Sekine & Elisabete Ranchhod (eds.), Journal **Linguisticae Investigationes**, Special Issue on Named Entity Recognition and Categorisation, LI 30:1, pp. 135-162. John Benjamins Publishing Company. ISSN 0378-4169.

NER and — Name variant merging

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 NewsExplorer name variants in 3 steps:

- Transliteration
- Normalisation
- Similarity measure

Name normalisation — Why? (1)

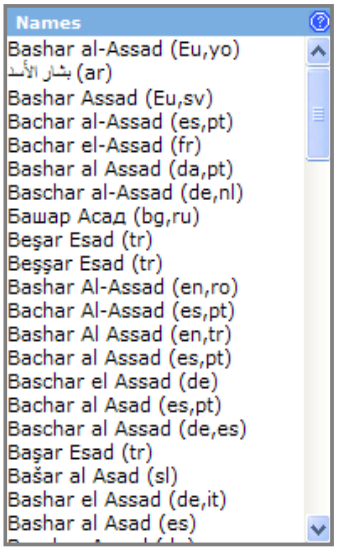
Transliteration rules depend on the target language, e.g.

Владимир Устинов (Russian)

- Vladimir Ustinov (English)
- Wladimir Ustinow (German)
- Vladimir Oustinov (French)

Various ways to represent the same sound:
sh, sch, ch, š, e.g.

- Bašar al Assad
- Baschar al Assad
- Bachar al Assad



Name normalisation — Why? (2)

Diacritics are often omitted, e.g.

- Wałęsa → Walesa
- Saïd → Said
- Schröder → Schroder
- Skarsgård → Skarsgard
- Jørgen → Jorgen

→ Levenshtein edit distance is large for naturally occurring word variants:

- “Rafik Harriri" vs. "Rafiq Hariri“ → 2
- “Rfk Hrr" vs. "Rafiq Hariri“ → 6

Name normalisation — How? ~ 30 rules

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
- …
- Remove vowels

→ Consonant signature

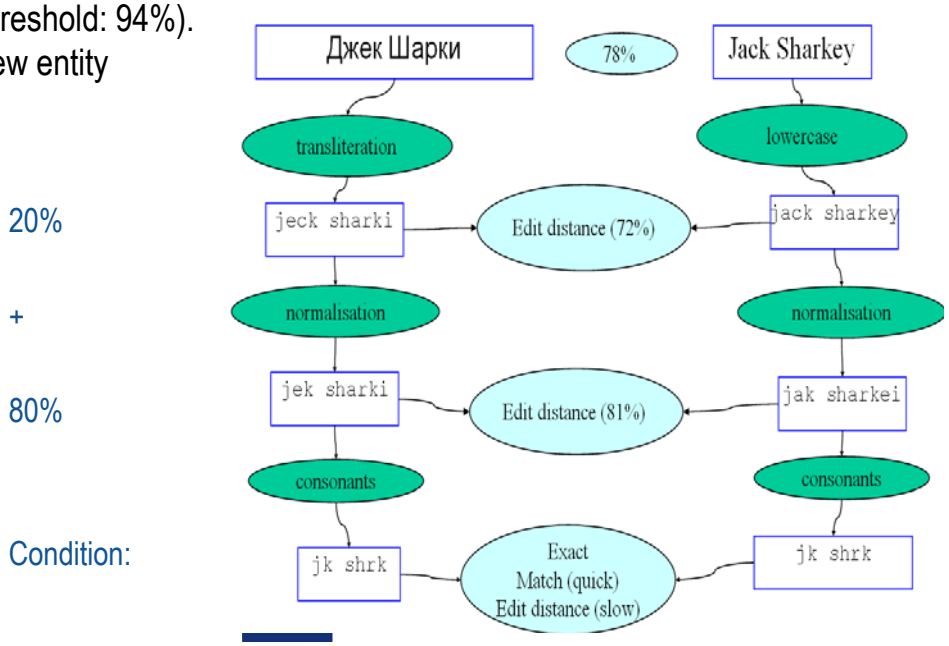
Malik al-Saïdoullaïev
→ Malik al-Saidoullaiev
→ Malik al-Saidoulaiev
→ Malik al-Saidulaiev
→ Malik Saidulaiev
→ … mlk sdlv

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

Merging variants for the same entity

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



Manual moderation of EMM name data

Process is fully automatic, but it can be useful to make changes manually.
Manual process only for frequent or important names (e.g. Nobel Prize winners):

- Name changes: (e.g. Cardinal Josef Ratzinger → Pope Benedict XVI)
- Correct NER mistakes (e.g. *Genius Report*, *Opfer von Diskriminierung*);
- Add new stop name parts (e.g. *Monday*, *Report*);
- Merge name variants with similarity below the threshold;
- Change the display name of an entity;
- Correct the entity type (PER, ORG, T, U, …);
- Launch Wikipedia mining process;
- …

Caveat: Name database contains errors!

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

Agenda

- EC-JRC: Who we are and what we do
- What is JRC-Names; What can it be used for
- Related work: other named entity (NE) resources
- How JRC-Names was produced
- Statistics on JRC-Names
- Current work: multilingual acronym recognition
- Further multilingual linguistic resources
- Summary

Entity 100006

Giovanni+Bosco
John+Bosco
Don+Bosco
· · + · · ·
Jean+Bosco
Joao+Bosco
Sveti+Janez+Bosco
Jan+Bosco
Yohannes+Bosco
Św+Jonas+Boskas
若望+鮑思高
Gioan+Bosco
Bosco+Szent+Janos
Iohannes+Bosco
若望鮑思高
Ioan+Bosco
Jan+Bosko
ヨハネボスコ
Don+Bosko
João+Bosco
Bosco+Szent+János
ヨハネ+ボスコ
Ivan+Bosco
若望·鮑思高
Juan+Bosco
Giovanni+Melchior+Bosco
นักบุญยอห์น+บอสโก
Johannes+Bosco
ヨハネ・ボスコ
Sant+Joan+Bosco
Иоанн+Боско
Sveti+Ivan+Bosco

Statistics on JRC-Names (1)

JRC-Names include names from the EMM database if any of the following hold:

- Found in 5 or more news clusters;
- Manually verified;
- Retrieved from Wikipedia;

Number of entries (status July 2011):

- 205,000 distinct names;
- 204,000 additional variants;
- ~3.2% names of organisations / events

Number of variants:

- 413 variants for *Muammar Gaddafi* (entity 262)
- 256 variants for *Mikhail Saakashvili* (entity 472)
- 246 variants for *Mahmoud Ahmadinejad* (entity 101358)

Grows by almost 1000 new names or variants per week.

Variant forms	Nº. of entities
1	63.76%
2	22.52%
3	5.31%
10 or more	3760 entities
50 or more	242 entities
100 or more	37 entities

Statistics on JRC-Names (2)

Number of scripts: 27

ISO15924	TEXT	Number Variants	Count Entities
Latn	Latin	1588622	1263969
Cyrl	Cyrillic	104107	88097
Arab	Arabic	17691	14513
Jpan	Japanese (Han+Hiragana+Katakana)	6995	6785
Hans	Han (Simplified variant)	4751	4512
Hebr	Hebrew	3811	3664
Kore	Korean (Hangul+Han)	2432	2354
Deva	Devanagari (Nagari)	1527	1043
Grek	Greek	1476	1410
Thai	Thai	1203	1140
Geor	Georgian (Mkhedruli)	1072	1021
Beng	Bengali	674	645
Taml	Tamil	639	618
Mlym	Malayalam	278	272
Armn	Armenian	195	188
Knda	Kannada	145	139
Telu	Telugu	128	126
Ethi	Ethiopic (Ge□ez)	112	108

Number of languages: ???

- News mentions names from around the world.
- Frequency does not reflect origin
  - *European Union* (10101) is most frequent entity in German, and second in English.
- It does not matter where a name like *Silvio Berlusconi* comes from.

(status July 2011)

Agenda

- EC-JRC: Who we are and what we do
- What is JRC-Names; What can it be used for
- Related work: other named entity (NE) resources
- How JRC-Names was produced
- Statistics on JRC-Names
- Current work: multilingual acronym recognition
- Further multilingual linguistic resources
- Summary

Entity 100006
Giovanni+Bosco
John+Bosco
Don+Bosco
. . + . . .
Jean+Bosco
Joao+Bosco
Sveti+Janez+Bosco
Jan+Bosco
Yohannes+Bosco
Św+Jonas+Boskas
若望+鮑思高
Gioan+Bosco
Bosco+Szent+Janos
Iohannes+Bosco
若望鮑思高
Ioan+Bosco
Jan+Bosko
ヨハネボスコ
Don+Bosko
João+Bosco
Bosco+Szent+János
ヨハネ+ボスコ
Ivan+Bosco
若望·鮑思高
Juan+Bosco
Giovanni+Melchior+Bosco
นักบุญยอห์น+บอสโก
Johannes+Bosco
ヨハネ・ボスコ
Sant+Joan+Bosco
Иоанн+Боско
Sveti+Ivan+Bosco

Current work — Acronym recognition

- Pairs of Long-Form (LF) and their Short-Form (SF)
  - e.g. … by the *International Monetary Fund (IMF)*.
- Method: search left-hand-side context of bracket
- Allows to recognise referring expressions of very different types, including:
  - Organisations (Center for Autism Research)
  - Programs (FP7)
  - Country names (Central African Republic)
  - Methods (Computer-assisted Reporting)
  - Medical terminology (Magnetic Resonance Imaging)
- We currently ignore other formats:
  - … by the *IMF (International Monetary Fund)*.
  - … while the *International Monetary Fund, IMF*, currently …
- Plan: distribute together with *JRC-Names.*

Ehrmann Maud, Leonida della Rocca, Ralf Steinberger & Hristo Tanev (submitted). **Acronym recognition and processing in 22 languages.**

CAR

Found in English text
capital adequacy ratio
Capital Adequate Ratio
Capital Adequacy Ration
Capital Adequacy Returns
Center for Autism Research
central African Republic
Certified Automotive Recycler Program
Commission for Aviation Regulation
Confederations of Africa Rugby
Cordilleral Administrative Region

Found in French text
Caisse Autonome des Retraites
capacité africaine contre les risques
Cellule d'Action Routière
Collectif d'artistes de reggae
Collectivité d'accueil régionale
Comité d'Action pour le Renouveau
Communauté d'agglomération de Rufisque

Found in German text
Centers for Automotive Research
Central African Republic
chimären Antigenrezeptoren
Computer Assisted Reporting

Found in Italian text
Cogenerazione ad Alto Rendimento
Computer Assisted Reporting
consumo annuo di riferimento

Acronym recognition — Statistics

Highly productive: we find 30-40,000 new acronyms per month (in 22 languages; Format LF (SF) only:

ISO	Language	% of text analysed	Acronym frequency	Re-use frequency	Usage once	Usage ≥10
Ca	Catalan	0.2%	21.2%	4.66	13.2%	1.29%
Cs	Czech	1.5%	8.9%	8.95	5.5%	1.18%
Da	Danish	3.3%	3.0%	9.92	5.6%	1.41%
De	German	12.7%	13.3%	9.09	6.2%	0.92%
En	English	**25.1%**	26.2%	7.51	7.8%	1.01%
Es	Spanish	11.9%	30.7%	11.64	5.0%	0.77%
Et	Estonian	0.9%	3.9%	3.76	16.7%	1.42%
Eu	Basque	**0.0%**	2.9%	**1.98**	**34.8%**	1.11%
Fi	Finnish	2.2%	**1.4%**	4.33	15.0%	1.59%
Fr	French	8.8%	28.8%	6.59	9.3%	1.04%
Hu	Hungarian	2.7%	9.5%	8.79	6.4%	0.96%
It	Italian	4.8%	3.2%	2.48	28.7%	1.22%
Lt	Lithuanian	1.0%	22.3%	10.43	5.0%	1.00%
Lv	Latvian	0.9%	**32.4%**	**14.67**	3.7%	0.78%
Nl	Dutch	4.2%	8.0%	7.25	8.2%	1.04%
No	Norwegian	1.4%	6.2%	5.41	11.3%	1.27%
Pl	Polish	2.5%	3.9%	4.65	12.0%	**1.58%**
Pt	Portuguese	4.9%	27.9%	13.13	**3.5%**	0.84%
Ro	Romanian	5.7%	13.5%	8.11	7.3%	0.97%
Sl	Slovene	1.1%	9.2%	6.16	9.0%	1.45%
Sv	Swedish	4.0%	2.4%	6.96	8.4%	1.52%
Sw	Swahili	0.1%	16.6%	4.32	17.9%	**0.74%**
TOTAL		**100%**	**18.6%**		**7.13%**	**0.95%**

CAR

Found in English text
capital adequacy ratio
Capital Adequate Ratio
Capital Adequacy Ration
Capital Adequacy Returns
Center for Autism Research
central African Republic
Certified Automotive Recycler Program
Commission for Aviation Regulation
Confederations of Africa Rugby
Cordilleral Administrative Region

Found in French text
Caisse Autonome des Retraites
capacité africaine contre les risques
Cellule d'Action Routière
Collectif d'artistes de reggae
Collectivité d'accueil régionale
Comité d'Action pour le Renouveau
Communauté d'agglomération de Rufisque

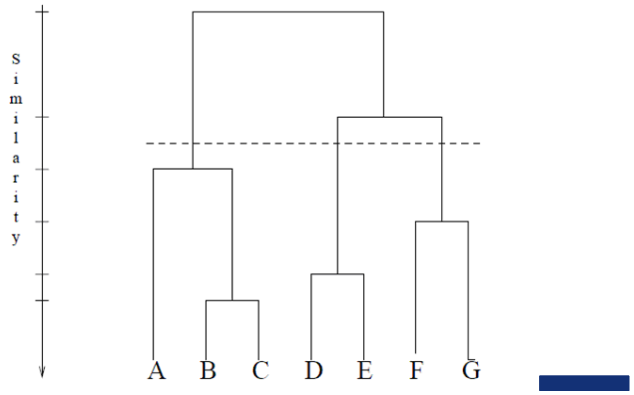
Found in German text
Centers for Automotive Research
Central African Republic
chimären Antigenrezeptoren
Computer Assisted Reporting

Found in Italian text
Cogenerazione ad Alto Rendimento
Computer Assisted Reporting
consumo annuo di riferimento

Acronyms — Identify variants

- In news, many spelling variants are used.
- Challenge: Identify related variants.
- Method: hierarchical clustering, using cosine.
- Similarity measure: normalised Levenshtein edit distance (letter insertions + deletions + substitutions / number of letters)
  - kitten → sitten (substitution of "s" for "k")
  - sitten → sittin (substitution of "i" for "e")
  - sittin → sitting (insertion of "g" at the end).

  Kitten → sitting: Similarity = 1 - 3/7 = 1-0.43 = 0.57



Agenzia internazionale per l'energia atomica (AIEA)
agenzia delle Nazioni Unite per l'energia atomica
Agenzia di controllo sul nucleare delle Nazioni Unite
Agenzia internazionale energia atomica
Agenzia Internazionale delEnergia Atomica
Agenzia internazionale dell'Onu per l'energia atomica
Agenzia internazionale delOnu per energia atomica
Agenzia internazionale Energia atomica
Agenzia Internazionale Onu per l'Energia Atomica
agenzia Internazionale per Energia Atomica
Agenzia internazionale per il nucleare
Agenzia Internazionale per la Sicurezza Nucleare
Agenzia internazionale per l'energia atomica Onu
Agenzia nucleare delOnu
Agenzia Onu per il Nucleare
Agenzia Onu sul nucleare
Agenzia per l'Energia atomica
Agenzia per l'energia nucleare Onu
all'Agenzia internazionale dell'energia atomica
all'Organizzazione iraniana dell'energia atomica
Atomic Energy Agency
Atomica delle Nazioni Unite
dell'Agenzia dell'Onu sul nucleare
dell'Agenzia delle Nazioni Unite per l'energia atomica
dell'Agenzia di controllo sul nucleare delle Nazioni Unite
dell'agenzia nazionale per l'energia atomica

Agenda

- EC-JRC: Who we are and what we do
- What is JRC-Names; What can it be used for
- Related work: other named entity (NE) resources
- How JRC-Names was produced
- Statistics on JRC-Names
- Current work: multilingual acronym recognition
- Further multilingual linguistic resources
- Summary

Entity 100006
Giovanni+Bosco
John+Bosco
Don+Bosco
...+...
Jean+Bosco
Joao+Bosco
Sveti+Janez+Bosco
Jan+Bosco
Yohannes+Bosco
Św+Jonas+Boskas
若望+鮑思高
Gioan+Bosco
Bosco+Szent+Janos
Iohannes+Bosco
若望鮑思高
Ioan+Bosco
Jan+Bosko
ヨハネボスコ
Don+Bosko
João+Bosco
Bosco+Szent+János
ヨハネ+ボスコ
Ivan+Bosco
若望·鮑思高
Juan+Bosco
Giovanni+Melchior+Bosco
นักบุญยอห์น+บอสโก
Johannes+Bosco
ヨハネ・ボスコ
Sant+Joan+Bosco
Иоанн+Боско
Sveti+Ivan+Bosco

Further multilingual linguistic resources

Parallel corpora (mostly sentence-aligned)

JRC-Acquis (2006): 1 billion word parallel corpus in 22 languages

DGT-Acquis (2012): 1 billion word parallel corpus in 22 languages

DCEP (Digital Corpus of the European Parliament) (forthcoming).
>1 billion word parallel corpus in 23 languages

Translation Memories

DGT-TM (since 2007): Translation Memory in 22 languages; up to 2 million segments; yearly updates (up to and including 2012)

ECDC-TM (2012): Translation Memory in 25 languages; 32K segments; *Public Health*.

EAC-TM (2013): Translation Memory in 26 languages; 78K segments; *Education & Culture*.

Document classification software

JEX (JRC Eurovoc Indexer) (2012): software to automatically label texts according to the thousands of categories of the Eurovoc thesaurus; 22 languages.

+ further smaller resources

All available for download from http://ipsc.jrc.ec.europa.eu/?id=61

Agenda

- EC-JRC: Who we are and what we do
- What is JRC-Names; What can it be used for
- Related work: other named entity (NE) resources
- How JRC-Names was produced
- Statistics on JRC-Names
- Current work: multilingual acronym recognition
- Further multilingual linguistic resources
- Summary

Entity 100006
Giovanni+Bosco
John+Bosco
Don+Bosco
...+...
Jean+Bosco
Joao+Bosco
Sveti+Janez+Bosco
Jan+Bosco
Yohannes+Bosco
Św+Jonas+Boskas
若望+鮑思高
Gioan+Bosco
Bosco+Szent+Janos
Iohannes+Bosco
若望鮑思高
Ioan+Bosco
Jan+Bosko
ヨハネボスコ
Don+Bosko
João+Bosco
Bosco+Szent+János
ヨハネ+ボスコ
Ivan+Bosco
若望·鮑思高
Juan+Bosco
Giovanni+Melchior+Bosco
นักบุญยอห์น+บอสโก
Johannes+Bosco
ヨハネ・ボスコ
Sant+Joan+Bosco
Иоанн+Боско
Sveti+Ivan+Bosco

JRC-Names

Summary

JRC-Names is a name variant resource.

- Name dictionary and software.
- Possible uses.
- How it was produced.
- Data-driven, not introspection.

Question: do we need dictionaries of names, acronyms, etc.?

- Might be useful for human translators (or consult http://emm.newsexplorer.eu instead?).

- From our own, data and application-oriented computational linguistics view: YES! E.g. to
  - Normalise name spelling variants in data bases;
  - Improve information retrieval;
  - Improve machine translation;
  - Develop multilingual Named Entity Recognition (NER) software;
  - And many more text mining applications.