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/

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

~ 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:
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

Monolingual PD work at the JRC

- 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)
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

---

**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

---

**Summary**

- A genome-wide association study (GWAS study) is an examination of genetic variation across the human genome, designed to identify genetic associations with observable traits, such as blood pressure or weight, or who some people get a disease or condition.
- These studies require large groups of participants with the disorder and age-matched unaffected individuals. After obtaining samples from an individual, the set of markers such as SNPs are scanned into computers. The computers search each participant’s genome for markers of genetic variation.
- Genetic variation explains the physical differences among people, such as eye color and blood group. Genetic variation also explains why some people are relatively rare disorders, such as cystic fibrosis and muscular dystrophy, or inherit an increased risk of common diseases such as cancer, heart disease and diabetes. Understanding how 0.1 percent of human genetic variation influences health and disease is one of medical science’s highest priorities.
Cross-lingual similarity

Entity names

Names

Bashar al-Assad (Eu,yo)
Bashar al-Assad (es,pt)
Bashar al-Assad (de,nl)
Bashar al Assad (da,sv)
Saawap Acazu (bg,ru)
Beșar Esad (tr)
Beşar Esad (fr,pt)
Bachar al Assad (es,pt)
Bachar al As (da,sv)
Bachar al Assad (es,sw)
Bacha al Assad (sl)
Bachar al Assad (de,es)
Bachar al As (es)
Bashir al-Assad (da,sv)

Key Titles and Phrases

syrian president (en - 1322)
president syrien (fr - 570)
presidente sirio (pt - 271)
presidente sirio (es - 205)
presidente siriano (it - 202)
president (de,sv - 961)
syrache président (nl - 133)
president (de - 446)
presidenten (de - 223)
siria (pt - 122)
prezydent (pl - 158)
staatschef (de - 92)
presidente (es,pt - 241)
president (fr - 249)
president (pl - 75)
siria (es - 51)
bağkani (tr - 209)
syrine (fr - 39)
predsednik (sl - 55)
statsminister (de - 35)
predsednik (sl - 51)

External resources

Image obtained automatically from Wikipedia

Read Wikipedia entry

Cross-lingual similarity

Entity names (2)

Names

Bashar al-Assad (Eu,yo)
Bashar al-Assad (es,pt)
Bashar al-Assad (de,nl)
Bashar al Assad (da,sv)
Saawap Acazu (bg,ru)
Beșar Esad (tr)
Beşar Esad (fr,pt)
Bachar al Assad (es,pt)
Bachar al As (da,sv)
Bachar al Assad (es,sw)
Bacha al Assad (sl)
Bachar al Assad (de,es)
Bachar al As (es)
Bashir al-Assad (da,sv)

Key Titles and Phrases

syrian president (en - 1322)
president syrien (fr - 570)
presidente sirio (pt - 271)
presidente sirio (es - 205)
presidente siriano (it - 202)
president (de,sv - 961)
syrache président (nl - 133)
president (de - 446)
presidenten (de - 223)
siria (pt - 122)
prezydent (pl - 158)
staatschef (de - 92)
presidente (es,pt - 241)
president (fr - 249)
president (pl - 75)
siria (es - 51)
bağkani (tr - 209)
syrine (fr - 39)
predsednik (sl - 55)
statsminister (de - 35)
predsednik (sl - 51)

Associated People

Kofi Annan (2563)
Bennett Miller (1306)
Hillary Rodham Clinton (1474)
Barack Obama (1754)
Karel Tevory Ingwersen (761)
Nabil Elaraby (887)
Vladimir Putin (926)
Muammar Gaddafi (805)
Ahmet Davutoglu (805)
Alan Jupp (785)
Abdul Rahman (710)
Walid Msalem (648)
Heikko Ahonen (243)
William Hague (591)
Catherine Ashton (511)
Reham Chahine (509)
Susan Kha (448)
Guido Westerwelle (436)
David Cameron (419)
Viktor Orbán (402)
François Hollande (297)
László Fabus (384)
Robin Hoods (375)
Mahmoud Ahmadinejad (372)
Nest Pillay (270)

News and organizations

Main Menu

News and organizations

Bashar Assad

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

Names

Bashar al-Assad (Eu,yo)
Bashar al-Assad (es,pt)
Bashar al-Assad (de,nl)
Bashar al Assad (da,sv)
Saawap Acazu (bg,ru)
Beșar Esad (tr)
Beşar Esad (fr,pt)
Bachar al Assad (es,pt)
Bachar al As (da,sv)
Bachar al Assad (es,sw)
Bacha al Assad (sl)
Bachar al Assad (de,es)
Bachar al As (es)
Bashir al-Assad (da,sv)

Key Titles and Phrases

syrian president (en - 1322)
president syrien (fr - 570)
presidente sirio (pt - 271)
presidente sirio (es - 205)
presidente siriano (it - 202)
president (de,sv - 961)
syrache président (nl - 133)
president (de - 446)
presidenten (de - 223)
siria (pt - 122)
prezydent (pl - 158)
staatschef (de - 92)
presidente (es,pt - 241)
president (fr - 249)
president (pl - 75)
siria (es - 51)
bağkani (tr - 209)
syrine (fr - 39)
predsednik (sl - 55)
statsminister (de - 35)
predsednik (sl - 51)

External resources

Image obtained automatically from Wikipedia

Read Wikipedia entry

Timeline

Timeline for 09-SEP-12

State TV reports 6 dead in Damascus 'terrorists' beat

Clintec tells Russia sanctions will end, but Congress may disagree

JHT SB-SEP-12

Russia named the conditions for a peace settlement in Syria

Associated People

Ali Akbarzadeh (10)
Mustafa al-Dabbagh (6)
Vladimir Petrovich Khudyev (1)
Vladimir Klyuyev (1)
Reham Chahine (1)

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

asesinato del exprimer ministro Rafic al-Hariri, que la oposición atribuyó

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

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

danjega libanonskega premiera Rafika Haririja. Libanonska opozicija si

möödumisele ekspeaminister Rafik al-Hariri surma põhjustanud pommipl

اغتيال رئيس الوزراء السابق رفيق الحريري بآياد يهودية وما حدث سابقا

Бывший премьер-министр Ливана Рафик Харири, который
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.

For example:

- Hamid Karzai
- Hamid Karzaï
- Hamid Karsai
- كرزاي
- حامد
- हामद करजई
- 卡尔扎伊

New name variants, including across scripts, will be recognised in new EMM articles.
• 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
Cross-lingual similarity

*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)

  + Achieved results are relatively good
  - Bilingual approach is restricted to a few languages

\[
\text{Language pairs} = N * (N - 1) / 2 \\
20 \text{ NewsExplorer languages} \rightarrow 190 \text{ language pairs (380 language pair directions)!}
\]
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

---

CL Document Similarity

Language-independent anchors

Language-independent features for multilingual document representation

No MT or bilingual dictionaries

20 languages

\[
\text{CLDS} = \alpha \cdot S_1 + \beta \cdot S_2 + \gamma \cdot S_3 + \delta \cdot S_4
\]
• **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.

<table>
<thead>
<tr>
<th>Similarity threshold</th>
<th>FR⁺</th>
<th>FR⁻</th>
<th>IT⁺</th>
<th>IT⁻</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 - 19%</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>20 - 29%</td>
<td>1</td>
<td>6</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>30 - 39%</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>40 - 49%</td>
<td>16</td>
<td>4</td>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>50 - 59%</td>
<td>19</td>
<td>1</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>60 - 100%</td>
<td>34</td>
<td>1</td>
<td>29</td>
<td>1</td>
</tr>
</tbody>
</table>

Accuracy at 30% threshold: R=99%, P=88%, R=97%, P=89%

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

---

**JRC EuroVoc Indexer JEX**

- JEX is multilingual multi-label classification software
- Using the controlled vocabulary from EuroVoc (>6,000 classes)
- 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
- JEX is freely downloadable from [http://langtech.jrc.ec.europa.eu/Eurovoc.html](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.
**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**

---

**JEX evaluation for 22 languages**

<table>
<thead>
<tr>
<th>Language</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>F1 dynamic rank</th>
<th>Categories collection</th>
<th>Average categories trained</th>
<th>Stop words used (+MW)</th>
<th>Total number of documents</th>
<th>Document length (words)</th>
<th>All categories trained</th>
</tr>
</thead>
<tbody>
<tr>
<td>BG</td>
<td>0.4619</td>
<td>0.5120</td>
<td>0.4857</td>
<td>0.4940</td>
<td>3780</td>
<td>2049.9</td>
<td>332</td>
<td>22696</td>
<td>786.96x2764.72</td>
<td>2147</td>
</tr>
<tr>
<td>CS</td>
<td>0.4669</td>
<td>0.5208</td>
<td>0.4935</td>
<td>0.4960</td>
<td>3691</td>
<td>2035.7</td>
<td>137</td>
<td>22630</td>
<td>890.66x3317.10</td>
<td>2129</td>
</tr>
<tr>
<td>DA</td>
<td>0.4747</td>
<td>0.5491</td>
<td>0.5092</td>
<td>0.5170</td>
<td>4226</td>
<td>2655</td>
<td>858</td>
<td>41727</td>
<td>561.87x1875.19</td>
<td>2752</td>
</tr>
<tr>
<td>DE</td>
<td>0.4732</td>
<td>0.5485</td>
<td>0.5081</td>
<td>0.5187</td>
<td>4230</td>
<td>2683.9</td>
<td>1793</td>
<td>41676</td>
<td>566.67x1945.44</td>
<td>2783</td>
</tr>
<tr>
<td>EL</td>
<td>0.4632</td>
<td>0.5369</td>
<td>0.4975</td>
<td>0.5118</td>
<td>4214</td>
<td>2486.4</td>
<td>105</td>
<td>41013</td>
<td>778.05x2379.45</td>
<td>2605</td>
</tr>
<tr>
<td>EN</td>
<td>0.4801</td>
<td>0.5547</td>
<td>0.5227</td>
<td>0.5229</td>
<td>4229</td>
<td>2324.1</td>
<td>1180</td>
<td>41989</td>
<td>663.33x2204.28</td>
<td>2688</td>
</tr>
<tr>
<td>ES</td>
<td>0.4801</td>
<td>0.5545</td>
<td>0.5227</td>
<td>0.5229</td>
<td>4221</td>
<td>2297</td>
<td>481 (+264)</td>
<td>41397</td>
<td>547.63x1819.14</td>
<td>2406</td>
</tr>
<tr>
<td>ET</td>
<td>0.4828</td>
<td>0.5358</td>
<td>0.5075</td>
<td>0.5139</td>
<td>3662</td>
<td>2047.8</td>
<td>1533</td>
<td>21989</td>
<td>652.22x2193.32</td>
<td>2120</td>
</tr>
<tr>
<td>FI</td>
<td>0.4654</td>
<td>0.5341</td>
<td>0.4974</td>
<td>0.5081</td>
<td>4103</td>
<td>2528.8</td>
<td>92</td>
<td>38293</td>
<td>756.70x2565.81</td>
<td>2634</td>
</tr>
<tr>
<td>FR</td>
<td>0.4776</td>
<td>0.5536</td>
<td>0.5227</td>
<td>0.5229</td>
<td>4234</td>
<td>2588.7</td>
<td>1180</td>
<td>41989</td>
<td>663.33x2204.28</td>
<td>2688</td>
</tr>
<tr>
<td>HU</td>
<td>0.5121</td>
<td>0.5654</td>
<td>0.5374</td>
<td>0.5444</td>
<td>3585</td>
<td>1688.5</td>
<td>1228 (+709)</td>
<td>20838</td>
<td>851.66x1977.14</td>
<td>1788</td>
</tr>
<tr>
<td>IT</td>
<td>0.4713</td>
<td>0.5468</td>
<td>0.5060</td>
<td>0.5151</td>
<td>4234</td>
<td>2584.4</td>
<td>219</td>
<td>41883</td>
<td>764.57x2808.31</td>
<td>2688</td>
</tr>
<tr>
<td>LT</td>
<td>0.4920</td>
<td>0.5454</td>
<td>0.5075</td>
<td>0.5239</td>
<td>3635</td>
<td>1945.7</td>
<td>1199</td>
<td>21505</td>
<td>644.53x2724.18</td>
<td>2046</td>
</tr>
<tr>
<td>LV</td>
<td>0.4659</td>
<td>0.5175</td>
<td>0.4904</td>
<td>0.4968</td>
<td>3690</td>
<td>2011</td>
<td>14</td>
<td>22803</td>
<td>894.59x3012.39</td>
<td>2106</td>
</tr>
<tr>
<td>MT</td>
<td>0.4200</td>
<td>0.4545</td>
<td>0.4366</td>
<td>0.4416</td>
<td>3584</td>
<td>1762.3</td>
<td>6</td>
<td>17858</td>
<td>1016.99x2655.11</td>
<td>1864</td>
</tr>
<tr>
<td>NL</td>
<td>0.4803</td>
<td>0.5562</td>
<td>0.5257</td>
<td>0.5257</td>
<td>4232</td>
<td>2610.2</td>
<td>1414</td>
<td>41816</td>
<td>581.94x1819.66</td>
<td>2713</td>
</tr>
<tr>
<td>PL</td>
<td>0.4794</td>
<td>0.5111</td>
<td>0.4930</td>
<td>0.5077</td>
<td>3648</td>
<td>1967.1</td>
<td>125</td>
<td>22404</td>
<td>841.81x2795.25</td>
<td>2066</td>
</tr>
<tr>
<td>PT</td>
<td>0.4756</td>
<td>0.5403</td>
<td>0.5098</td>
<td>0.5237</td>
<td>4209</td>
<td>2560.6</td>
<td>1152</td>
<td>41142</td>
<td>700.46x2181.01</td>
<td>2663</td>
</tr>
<tr>
<td>RO</td>
<td>0.4550</td>
<td>0.5043</td>
<td>0.4784</td>
<td>0.4817</td>
<td>3887</td>
<td>2109.3</td>
<td>1594 (+48)</td>
<td>25023</td>
<td>994.17x3063.16</td>
<td>2206</td>
</tr>
<tr>
<td>SK</td>
<td>0.4705</td>
<td>0.5204</td>
<td>0.4942</td>
<td>0.4995</td>
<td>3645</td>
<td>1938.4</td>
<td>364</td>
<td>21406</td>
<td>872.53x3241.50</td>
<td>2050</td>
</tr>
<tr>
<td>SL</td>
<td>0.4840</td>
<td>0.5341</td>
<td>0.5078</td>
<td>0.5205</td>
<td>3645</td>
<td>2013.1</td>
<td>2068</td>
<td>22289</td>
<td>627.56x2698.38</td>
<td>2119</td>
</tr>
<tr>
<td>SV</td>
<td>0.4787</td>
<td>0.5472</td>
<td>0.5107</td>
<td>0.5194</td>
<td>4109</td>
<td>2546.4</td>
<td>1093</td>
<td>38198</td>
<td>609.82x2655.26</td>
<td>2649</td>
</tr>
</tbody>
</table>
• 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


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

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

Precision at rank 1.
• Setting a threshold, juggling Precision and Recall

<table>
<thead>
<tr>
<th>Test bed</th>
<th>Average similarity</th>
<th>Threshold</th>
<th>Recall</th>
<th>Noise (1-Precision)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set T1 (820)</td>
<td>0.82</td>
<td>0.70</td>
<td>90%</td>
<td>2.2%</td>
</tr>
<tr>
<td>Set T2 (795)</td>
<td>0.79</td>
<td>0.70</td>
<td>76.5%</td>
<td>5%</td>
</tr>
</tbody>
</table>

• Searching for a translation where there is none:
  • Searching in T2 for documents of T1
    $\rightarrow$ 4.15% noise

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

---

**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
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.

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
Freely available resources

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

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