

# Joint Research Centre

freely available at  
<http://langtech.jrc.it/JRC-Acquis/>

# The JRC-Acquis

## A Multilingual Aligned Parallel Corpus with 22 Languages

### What is the JRC-Acquis?

- A parallel text corpus in 22 official EU languages.
- With **paragraph alignments** for all 231 language pairs.
- The *Acquis Communautaire* (AC) is the body of common rights and obligations which bind all the Member States together within the European Union (EU).
- By definition, documents exist in **all official EU languages (Romanian and Bulgarian included)**. Croatian, Turkish, etc. translations are in preparation.
- It contains: documents on political objectives, treaties, declarations, resolutions, agreements, EU legislation, etc.
- The *JRC-Acquis* is a selection of those EU documents that are available (a) on the web and (b) in at least 10 languages, of which (c) at least three new EU languages.

### What is so special about the JRC-Acquis?

- It is the **biggest parallel corpus in so many languages** (almost 50 Million words per language, over 1 Billion words in total).
- It is available in **22 languages**.
- Pair-wise alignment for all 231 language pairs (including **rare language combinations** like Bulgarian-Estonian, etc.
- It is manually classified according to EUROVOC subject domains.
- It is **free** for download for non-commercial use.

**Unique!**

### Size of the JRC-Acquis (version 3)

Language ISO code	N° of texts	Text body			Signatures Total N° words	Annexes Total N° words	Total N° words (text + signatures + annexes)
		Total N° words	Total N° characters	Average n° words			
bg	10956	15208341	98288769	1388.13	2052773	12885853	30146967
cs	21438	22843279	148972581	1065.55	7225300	16763733	46832312
da	23624	31489627	213468135	1331.68	2629786	18856213	50944826
de	23541	32059892	232748675	1361.87	2542149	16327611	50929652
el	23184	36453749	239583543	1572.37	2973674	16459680	56807003
en	23545	34588383	210692059	1469.03	3198766	17750761	55537910
es	23573	38926161	238016756	1651.3	3490204	19176243	62132688
et	23541	24621625	192700704	1046.9	1136051	14995748	40953824
fr	23284	24833912	212178954	1058.67	2677798	12647371	40107981
fi	23527	39100499	234758290	1654.91	3021013	19978920	62100432
hu	22801	28602380	213804614	1264.44	2929488	14066496	46188364
it	23472	36764670	230677013	1523.72	3120797	18331536	57217002
lt	23379	26937773	199438268	1152.22	2436585	15010484	44392842
lv	22906	27592514	196452051	1204.6	1673124	15437969	44073607
mt	10545	20926909	128906748	1984.53	1336042	15620611	37883562
nl	23564	36285161	231963539	1496.57	3039680	18467115	56771856
pl	23478	29713003	214464026	1265.57	2513441	17027393	49253537
pt	23505	37221658	227499418	1533.56	3034308	19350227	59696203
ro	23573	9186947	69537301	1397.68	514296	11185842	20887085
sk	21943	26792637	178920434	1221.01	3227859	16190546	46211036
sl	20642	27702305	178651767	1342.04	3103193	16837717	47643215
sv	20243	29433037	199004401	1453.99	2575771	14966384	46974192
Total	463364	635283572	4282728446	1385.88	60251991	357770252	1053305415

### What can I use this parallel corpus for?

- Train automatic systems for Statistical Machine Translation.
- Produce multilingual lexical or semantic resources such as dictionaries or ontologies.
- Train and test software for multilingual information extraction and topic segmentation.
- Automatic translation consistency checking.
- Training of multilingual subject domain classifiers.
- Test and benchmark alignment software (sentences, words, etc.), across a larger variety of language pairs.
- Useful for all types of cross-lingual research.

### Alignment for 231 language pairs

- Paragraph level alignment (can be one sentence, part of a sentence, or more than one sentence).
- Aligned pair-wise for all 231 language pairs!** Using Vanilla (Gale & Church, 1993).
- For the documents of JRC-Acquis version 2.2, an additional alignment with HunAlign (Varga et al., 2005) is available.

### Sentence in 20 languages Two terms highlighted

en: <P>(11) The measures provided for in this Regulation are in accordance with the opinion of the **Management Committee for Pigeat**.</P>  
 cs: <P>(11) opatření tohoto nařízení jsou v souladu se stanoviskem **Řídícího výboru pro vepřové maso**.</P>  
 de: <P>(11) Die in dieser Verordnung festgesetzte Voranstaltungen erübrigen sich im Hinblick auf die Stellungnahme des **Fürvaltningskommitté for Svinekad**.</P>  
 de: <P>(11) Die in dieser Verordnung vorgesehenen Maßnahmen entsprechen der Stellungnahme des **Verwaltungsausschusses für Schweinefleisch**.</P>  
 el: <P>(11) Τα μέτρα που προβλέπονται στην παρούσα κανονισμός είναι σύμφωνα με τη γνώμη της **επιτροπής διαχείρισης χοιρινού κρέατος**.</P>  
 es: <P>(11) Las medidas previstas en el presente Reglamento se ajustan al dictamen del **Comité de gestión de la carne de porcino**.</P>  
 et: <P>(11) Käesoleva määrusega ettenähtud meetmed on kooskõlas **sealiha turu korralduskomitee** arvamusega.</P>  
 fi: <P>(11) Tässä asetuksessa säädytety toimenpiteet ovat **sianlihan hallintokomitean** lausunnon mukaiset.</P>  
 fr: <P>(11) Les mesures prévues par le présent règlement sont conformes à l'avis du **comité de gestion de la viande de porc**.</P>  
 hu: <P>(11) Az e rendeletben előírt rendelkezések összhangban vannak a **Sertéshúsiaki Irányítóbizottság** véleményével.</P>  
 it: <P>(11) Le misure previste dal presente regolamento sono conformi al parere del **comitato di gestione per le carni suine**.</P>  
 lt: <P>(11) Kadangi Šiame reglamente numatytos priemonės atitinka **Kaolienuos vadybos komiteto** nuomonę.</P>  
 lv: <P>(11) Šajā regulā paredzētie pasākumi ir saskaņā ar Cūkgaļas vadības komitejas atzinumu.</P>  
 mt: <P>(11) Il-miżuri li għalihom hemm provvediment f'dan ir-Regolament huma bi qbil ma' l-opinjoni tal-Kumitat ta' Gestiġni dwar il-Laham tal-Majjal.</P>  
 nl: <P>(11) De in deze verordening vervatte maatregelen zijn in overeenstemming met het advies van het **Comité van beheer voor varkensvlees**.</P>  
 pl: <P>(11) Środki przewidziane w niniejszym rozporządzeniu są zgodne z opinią **Komitetu Zarządzającego ds. Wieprzowiny**.</P>  
 pt: <P>(11) As medidas previstas no presente regulamento estão em conformidade com o parecer do **Comité de Gestão da Carne de Suíno**.</P>  
 sk: <P>(11) Opatrenia ustanovené v tomto nariadení sú v súlade so stanoviskom **Riadiaceho výboru pre bravčové mäso**.</P>  
 sv: <P>(11) UKrepi, dölöceni s to uredbu, so v skladu z mnenjem **Upravjalnega odbora za prašičje meso**.</P>  
 sv: <P>(11) De åtgärder som föreskrivs i denna förordning är förenliga med yttrandet från **Förvaltningskommittén för griskött**.</P>

### EUROVOC subject domain classification

- Documents are manually classified (multi-label, multi-domain)
- eurovoc** (<http://eurovoc.europa.eu/>):
- Translated 1-to-1 into more than 20 languages.
- Over 6000 hierarchically organised descriptors.
- Used by many parliaments etc. throughout Europe.
- Uses:
  - To derive domain-specific terminology.
  - As training and test data for multi-label classification tasks.
  - To create language-independent multi-faceted domain vectors for documents → cross-lingual linking of texts (e.g. <http://press.jrc.it/NewsExplorer>)

Some of the main subject domain classes of the JRC-Acquis:

Public Health	Veterinary Inspection
Environmental Protection	Animal Disease
Anti-Dumping Legislation	Merger Control
Control of State Aid	Tariff quota
Fruit Vegetable	Butter
White Sugar	Agricultural Product
Common Organisation of Markets	Accession to the EU
Import	Italy
Air Transport	Marketing Standard
Appointment of Staff	Exchange Rate

### Format of the corpus

- TEI P4-compliant XML format, in UTF-8.
- Cleaned of most noisy headers and footers; appendices marked up.
- Modular: download the languages you are interested in.
- Perl tool (included) lets you produce a corpus of all aligned sentences for the language pair of your choice.

### Contributors



R. Steinberger, C. Ignat, M. Kolar, A. Widiger, B. Pouliquen, (EU)  
 D. Tufiş, A. Ceausu, R. Ion, D. Stefanescu (Romania)  
 T. Erjavec (Slovenia)  
 D. Varga (Hungary)

