European Union Language Resources in Sketch Engine

Vít Baisa, Jan Michelfeit, Marek Medveď, Miloš Jakubíček
Masaryk University, Czech Republic, Brno
Lexical Computing Ltd, United Kingdom, Brighton
{vit.baisa,jan.michelfeit,marek.medved,milos.jakubicek}@sketchengine.co.uk

Abstract

Several parallel corpora built from European Union language resources are presented here. They were processed by state-of-the-art tools and made available for researchers in the Sketch Engine corpus management system. A completely new resource is introduced: EUR-Lex corpus, being one of the largest parallel corpus available at the moment, containing 840 million tokens of English and having the largest language pair (English-French) with more than 25 million aligned segments (paragraphs).

Keywords: JRC-Acquis, DCEP, DGT-TM, Europarl, EUR-Lex, Sketch Engine, parallel corpus, word sketch, parallel concordance

1. Introduction

The European Union is producing a large amount of valuable multilingual textual data every day. To be able to use it in applications, for text analysis, terminology extraction, full text search etc., it must be downloaded, converted into plain text, processed with suitable tools, aligned on sentence level and finally made available to researchers in some standard format. In this paper we describe our experience with using several resources built from European Union's (EU) multilingual resources, namely DCEP (Hajlaoui et al., 2014), DGT-TM (Steinberger et al., 2013) and Europarl (Koehn, 2005).

We also describe a new multilingual "EUR-Lex corpus" containing more than 840 million tokens of English. To our knowledge, it is currently the largest parallel corpus built from European language resources. The corpus was downloaded from the official website of EUR-Lex[1] which provides an access to up-to-date legal documents published by European Commission, European Parliament, national courts, Council of the European Union and other European institutions. The majority of recently added documents is translated into all official languages of EU making it a huge multilingual language resource.

Corpus	Tokens	Types	L	Format
JRC-Acquis	55,537,910	N/A	22	XML
DCEP	118,046,857	513,000	23	TXT
DGT-TM	74,365,007	342,340	24	TMX
Europarl	60,741,877	139,217	21	XML
EUR-Lex	839,745,466	2,416,841	24	various

Table 1: Comparison of various EU corpora.

All mentioned corpora are available for language researchers through the Sketch Engine corpus management system (Kilgarriff et al., 2014). EUR-Lex corpus is released in the form of gzipped archives containing a) documents with meta information in a flat XML format and b) alignment files for all language pairs. The whole gzipped dataset is over 40 GB.[2]

Table 1 compares the mentioned language resources. JRC-Acquis 3.0 figures[3] are there for comparison. "Tokens" is the number of tokens (words, numbers and punctuation) in the English parts of the corpora. "Types" is the number of unique English word forms, i.e. the size of English lexicons, "L" column contains the number of languages included and "Format" states in which form the source data is available.

Language	Since	ACQ	CEP	DGT	EUR	LEX
Dutch	1958	35	96	63	60	777
French	1958	39	116	47	67	878
German	1958	32	98	58	55	732
Italian	1958	36	103	66	59	807
Danish	1973	31	88	59	56	731
English	1973	35	118	74	61	840
Greek	1981	36	100	64	44	775
Portuguese	1986	37	99	66	61	793
Spanish	1986	39	106	69	61	831
Finnish	1995	25	72	47	41	558
Swedish	1995	29	86	55	52	640
Czech	2004	23	51	57	15	500
Estonian	2004	25	43	46	13	437
Hungarian	2004	29	50	55	15	500
Latvian	2004	28	48	54	14	491
Lithuanian	2004	27	47	52	14	476
Maltese	2004	21	46	30	—	466
Polish	2004	30	51	58	15	511
Slovak	2004	27	50	56	15	495
Slovenian	2004	28	50	57	15	509
Bulgarian	2007	16	41	33	11	457
Irish	2007	—	2	1	—	37
Romanian	2007	9	42	33	11	462
Croatian	2013	—	—	5	—	156

Table 2: Representation of languages (millions of tokens).

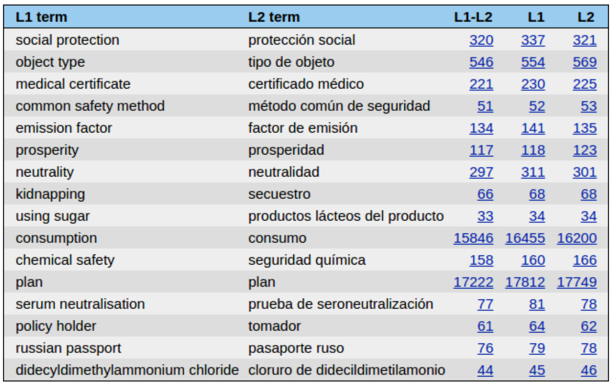
Table 2 contains an overview of language representation in the corpora in millions of tokens per language. The second column states a year when a particular language became an official language of European Union—it usually corresponds to the amount of documents in the particular lan-

[1] `http://eur-lex.europa.eu`

[2] To obtain the data, contact us or follow the instructions at `https://www.sketchengine.co.uk/eur-lex`

[3] `https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis`

guage and the table is sorted by this column. ACQ stands for JRC-Acquis 3.0, CEP for The Digital Corpus of the European Parliament, EUR for Europarl and LEX for EUR-Lex corpus.



L1 term	L2 term	L1-L2	L1	L2
social protection	protección social	320	337	321
object type	tipo de objeto	546	554	569
medical certificate	certificado médico	221	230	225
common safety method	método común de seguridad	51	52	53
emission factor	factor de emisión	134	141	135
prosperity	prosperidad	117	118	123
neutrality	neutralidad	297	311	301
kidnapping	secuestro	66	68	68
using sugar	productos lácteos del producto	33	34	34
consumption	consumo	15846	16455	16200
chemical safety	seguridad química	158	160	166
plan	plan	17222	17812	17749
serum neutralisation	prueba de seroneutralización	77	81	78
policy holder	tomador	61	64	62
russian passport	pasaporte ruso	76	79	78
didecyldimethylammonium chloride	cloruro de didecildimetilamonio	44	45	46

Figure 1: Bilingual terminology candidates extracted from DGT-TM English-Spanish.

2. DCEP

The Digital Corpus of the European Parliament (DCEP) (Hajlaoui et al., 2014) is a collection of documents published on the European Parliament's official website[4]. This corpus includes a variety of document types, from press release to session and legislative documents related to European Parliament's activities and bodies. The latest version contains documents produced in 2001–2012. Since the original alignments contained a lot of errors and the sentences were wrongly segmented, we created a new alignment. Instead of HunAlign (Varga et al., 2007) aligner we used GaChalign[5] algorithm (implementation of Gale-Church sentence aligner (Gale and Church, 1993)).

The data has been processed automatically by Sketch Engine: plain text data has been tokenized with unitok (Michelfeit et al., 2014) and tagged with various tools: TreeTagger (Schmid, 1995), Hunpos (Halácsy et al., 2007), Freeling (Carreras et al., 2004). Further processing involved collocation pattern extraction, terminology extraction, distributional thesaurus computation and other specific processing which is available in Sketch Engine for many languages (Kilgarriff et al., 2014).

3. DGT-TM

The European Commission's Directorate-General for Translation, in cooperation with the European Commission's Joint Research Centre, have created a freely available translation memory DGT-TM (Steinberger et al., 2013). The DGT-TM is stored in TMX files with segments aligned in 231 language pairs.

We have processed DGT-TM with Sketch Engine: it supports TMX import, we just merged all the original TMX files and let Sketch Engine extract the aligned segments, tokenize and PoS tag the texts. See Figure 2 for an example of parallel collocation functionality in Sketch Engine.

4. Europarl

The Europarl parallel corpus is a well-known resource (Koehn, 2005). It is a collection of sentence-aligned texts in 21 languages extracted from the proceedings of the European Parliament. It stands out among the other corpora provided by the EU, which contain mostly legal documents. Its primary goal is to aid statistical machine translation systems. The authors of the corpus have detected sentence boundaries in the raw transcripts and aligned the sentences using a tool based on the Church and Gale algorithm. (Gale and Church, 1993).

The Europarl corpus has been also incorporated into the OPUS project, a collection of publicly available parallel corpora (Tiedemann, 2009). Thanks to this, the sentence alignment data is available from the OPUS website in XCES format, which can be easily translated into the format used internally by Sketch Engine (pairs of structure IDs, here sentence IDs). See Figure 3 for an example of full text parallel search in Sketch Engine using Europarl corpus. All the text for each of the 21 languages was processed by the most up-to-date (at the time of compilation) processing chain for each respective language—including tokenization (Michelfeit et al., 2014), PoS tagging where available, but excluding sentence boundary detection, which was taken directly from Europarl data. Each of the resulting 21 corpora is therefore compatible for use as a reference corpus for other corpora in Sketch Engine (including user-created corpora) of the same language. The same holds for DCEP and DGT corpora. A reference corpus is used for comparison with a focus corpus for extraction of keywords and terminology. Bilingual terminology (Baisa et al., 2015) can be also extracted, see Figure 1.

All of the Europarl corpora are aligned to each other, giving us a total of 210 language pairs. Each pair of corpora can be exploited to extract a statistical dictionary of words and lemmas (where available), or even term candidates. Due to the nature of the texts, the vocabulary used is relatively broad, while the quality of the data is far better than other bigger, web-based corpora. This makes Europarl an invaluable resource for the creation of statistical dictionaries and building translation models for statistical machine translation systems.

5. EUR-Lex corpus

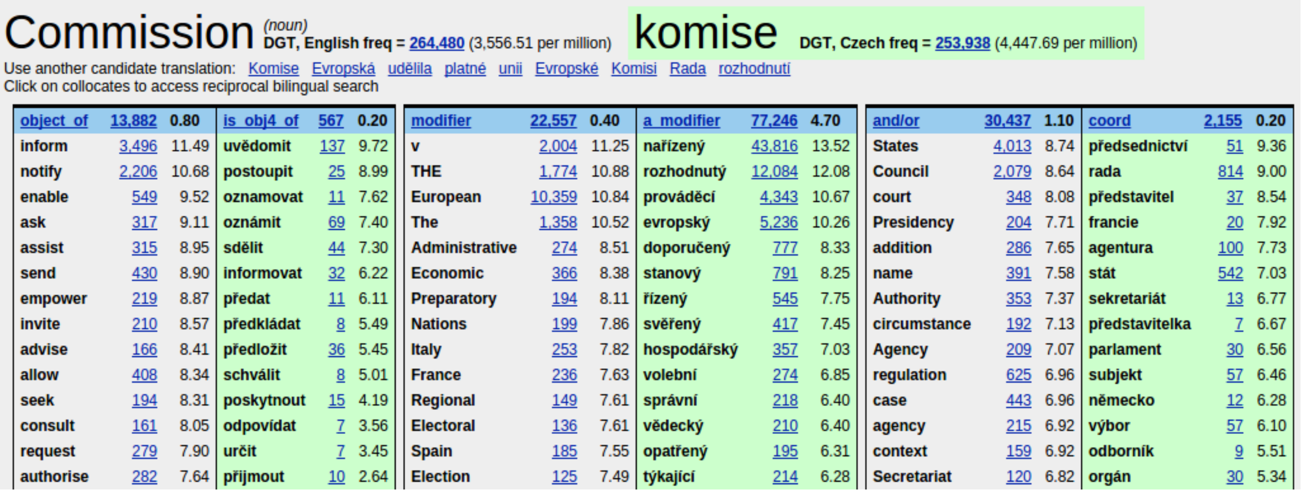
EUR-Lex is an official on-line resource providing access to 1) the Official Journal of the European Union, 2) EU law (EU treaties, directives, regulations, decisions, consolidated legislation, etc.), 3) preparatory acts (legislative proposals, reports, green and white papers, etc.), 4) EU case-law (judgements, orders, etc.), 5) international agreements, 6) EFTA documents and 7) other public documents dating back to 1950s in 24 official EU languages. The EUR-Lex website allows querying its database in which each document has meta data ranging from unique IDs (cellar and CELEX[6] numbering), dates of documents, official publication and revision dates, Eurovoc[7] terms, authors (an agent, a state) of a document, type of a document etc.

[4] http://www.europarl.europa.eu/

[5] https://github.com/alvations/gachalign

[6] http://eur-lex.europa.eu/content/help/faq/intro.html#help10

[7] http://eurovoc.europa.eu/

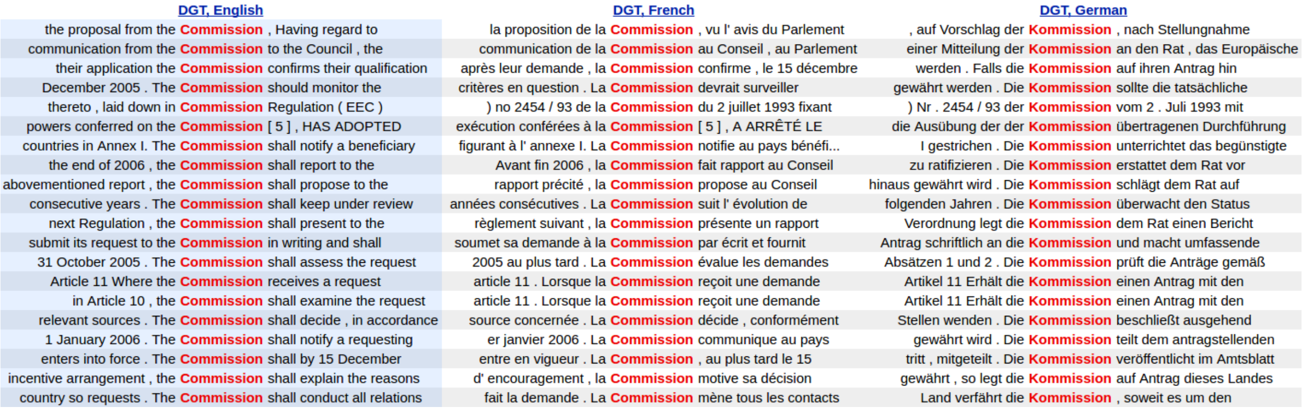


Commission (noun) DGT, English freq = 264,480 (3,556.51 per million) komise DGT, Czech freq = 253,938 (4,447.69 per million)

Use another candidate translation: Komise Evropská udělila platné unii Evropské Komisi Rada rozhodnutí
Click on collocates to access reciprocal bilingual search

object_of	13,882	0.80	is_obj4_of	567	0.20	modifier	22,557	0.40	a_modifier	77,246	4.70	and/or	30,437	1.10	coord	2,155	0.20
inform	3,496	11.49	uvědomit	137	9.72	v	2,004	11.25	nařízený	43,816	13.52	States	4,013	8.74	předsednictví	51	9.36
notify	2,206	10.68	postoupit	25	8.99	THE	1,774	10.88	rozhodnutý	12,084	12.08	Council	2,079	8.64	rada	814	9.00
enable	549	9.52	oznamovat	11	7.62	European	10,359	10.84	prováděcí	4,343	10.67	court	348	8.08	představitel	37	8.54
ask	317	9.11	oznámit	69	7.40	The	1,358	10.52	evropský	5,236	10.26	Presidency	204	7.71	francie	20	7.92
assist	315	8.95	sdělit	44	7.30	Administrative	274	8.51	doporučený	777	8.33	addition	286	7.65	agentura	100	7.73
send	430	8.90	informovat	32	6.22	Economic	366	8.38	stanový	791	8.25	name	391	7.58	stát	542	7.03
empower	219	8.87	předat	11	6.11	Preparatory	194	8.11	řízený	545	7.75	Authority	353	7.37	sekretariát	13	6.77
invite	210	8.57	předkládat	8	5.49	Nations	199	7.86	svěřený	417	7.45	circumstance	192	7.13	představitelka	7	6.67
advise	166	8.41	předložit	36	5.45	Italy	253	7.82	hospodářský	357	7.03	Agency	209	7.07	parlament	30	6.56
allow	408	8.34	schválit	8	5.01	France	236	7.63	volební	274	6.85	regulation	625	6.96	subjekt	57	6.46
seek	194	8.31	poskytnout	15	4.19	Regional	149	7.61	správní	218	6.40	case	443	6.96	německo	12	6.28
consult	161	8.05	odpovídat	7	3.56	Electoral	136	7.61	vědecký	210	6.40	agency	215	6.92	výbor	57	6.10
request	279	7.90	určit	7	3.45	Spain	185	7.55	opatřený	195	6.31	context	159	6.92	odborník	9	5.51
authorise	282	7.64	přijmout	10	2.64	Election	125	7.49	týkající	214	6.28	Secretariat	120	6.82	orgán	30	5.34

Figure 2: Parallel collocation candidates for English "Commission" and Czech equivalent "komise" derived from DGT-English and DGT-Czech corpora in Sketch Engine. The joint grey and green columns correspond to a grammar relation (object_of, modifier and coordination) in which the collocation candidates occur in data. The collocates in green columns are usually translation equivalents of the collocates in joint grey columns. E.g. inform—informovat, Electoral—volební, Presidency—předsednictví, etc.



DGT, English	DGT, French	DGT, German
the proposal from the Commission , Having regard to	la proposition de la Commission , vu l' avis du Parlement	, auf Vorschlag der Kommission , nach Stellungnahme
communication from the Commission to the Council , the	communication de la Commission au Conseil , au Parlement	einer Mitteilung der Kommission an den Rat , das Europäische
their application the Commission confirms their qualification	après leur demande , la Commission confirme , le 15 décembre	werden . Falls die Kommission auf ihren Antrag hin
December 2005 . The Commission should monitor the	critères en question . La Commission devrait surveiller	gewährt werden . Die Kommission sollte die tatsächliche
thereto , laid down in Commission Regulation (EEC)	) no 2454 / 93 de la Commission du 2 juillet 1993 fixant	) Nr . 2454 / 93 der Kommission vom 2 . Juli 1993 mit
powers conferred on the Commission [5] , HAS ADOPTED	exécution conférées à la Commission [5] , A ARRÊTÉ LE	die Ausübung der der Kommission übertragenen Durchführung
countries in Annex I. The Commission shall notify a beneficiary	figurant à l' annexe I. La Commission notifie au pays bénéfi...	I gestrichen . Die Kommission unterrichtet das begünstigte
the end of 2006 , the Commission shall report to the	Avant fin 2006 , la Commission fait rapport au Conseil	zu ratifizieren . Die Kommission erstattet dem Rat vor
abovementioned report , the Commission shall propose to the	rapport précité , la Commission propose au Conseil	hinaus gewährt wird . Die Kommission schlägt dem Rat auf
consecutive years . The Commission shall keep under review	années consécutives . La Commission suit l' évolution de	folgenden Jahren . Die Kommission überwacht den Status
next Regulation , the Commission shall present to the	règlement suivant , la Commission présente un rapport	Verordnung legt die Kommission dem Rat einen Bericht
submit its request to the Commission in writing and shall	soumet sa demande à la Commission par écrit et fournit	Antrag schriftlich an die Kommission und macht umfassende
31 October 2005 . The Commission shall assess the request	2005 au plus tard . La Commission évalue les demandes	Absätzen 1 und 2 . Die Kommission prüft die Anträge gemäß
Article 11 Where the Commission receives a request	article 11 . Lorsque la Commission reçoit une demande	Artikel 11 Erhält die Kommission einen Antrag mit den
in Article 10 , the Commission shall examine the request	article 11 . Lorsque la Commission reçoit une demande	Artikel 11 Erhält die Kommission einen Antrag mit den
relevant sources . The Commission shall decide , in accordance	source concernée . La Commission décide , conformément	Stellen wenden . Die Kommission beschließt ausgehend
1 January 2006 . The Commission shall notify a requesting	er janvier 2006 . La Commission communique au pays	gewährt wird . Die Kommission teilt dem antragstellenden
enters into force . The Commission shall by 15 December	entre en vigueur . La Commission , au plus tard le 15	tritt , mitgeteilt . Die Kommission veröffentlicht im Amtsblatt
incentive arrangement , the Commission shall explain the reasons	d' encouragement , la Commission motive sa décision	gewährt , so legt die Kommission auf Antrag dieses Landes
country so requests . The Commission shall conduct all relations	fait la demande . La Commission mène tous les contacts	Land verfährt die Kommission , soweit es um den

Figure 3: Parallel search in Sketch Engine for English Commission, French Commission and German Kommission, DGT.

To get all documents we first had to query EUR-Lex for meta data year by year as the list of all documents in EUR-Lex is not available. From the meta data, a list of all available documents with CELEX numbers was retrieved (with all its language variants) and then all the documents were downloaded: only documents in HTML format have been downloaded, yielding almost 7 million documents in 26 languages.[8] According to the statistics[9] there are more PDF documents than HTML documents but we decided to download only HTML in the first phase as HTML files are easier for further processing.

We have exploited the fact that EUR-Lex database contains HTML documents split into fine-grained paragraphs and these paragraphs mostly correspond to each other in different languages. This can be seen in the parallel view on the EUR-Lex website.[10] Sometimes, the count of paragraphs is inconsistent in some language mutations, so we have corrected these using a modified Gale-Church algorithm.[5]

The resulting corpus has 3.9 million documents. Figure 4 shows size of aligned documents. The largest language pair English-French has 25,211,093 aligned paragraphs. All data from JRC Acquis corpus (Steinberger et al., 2006) should be included in EUR-Lex corpus.

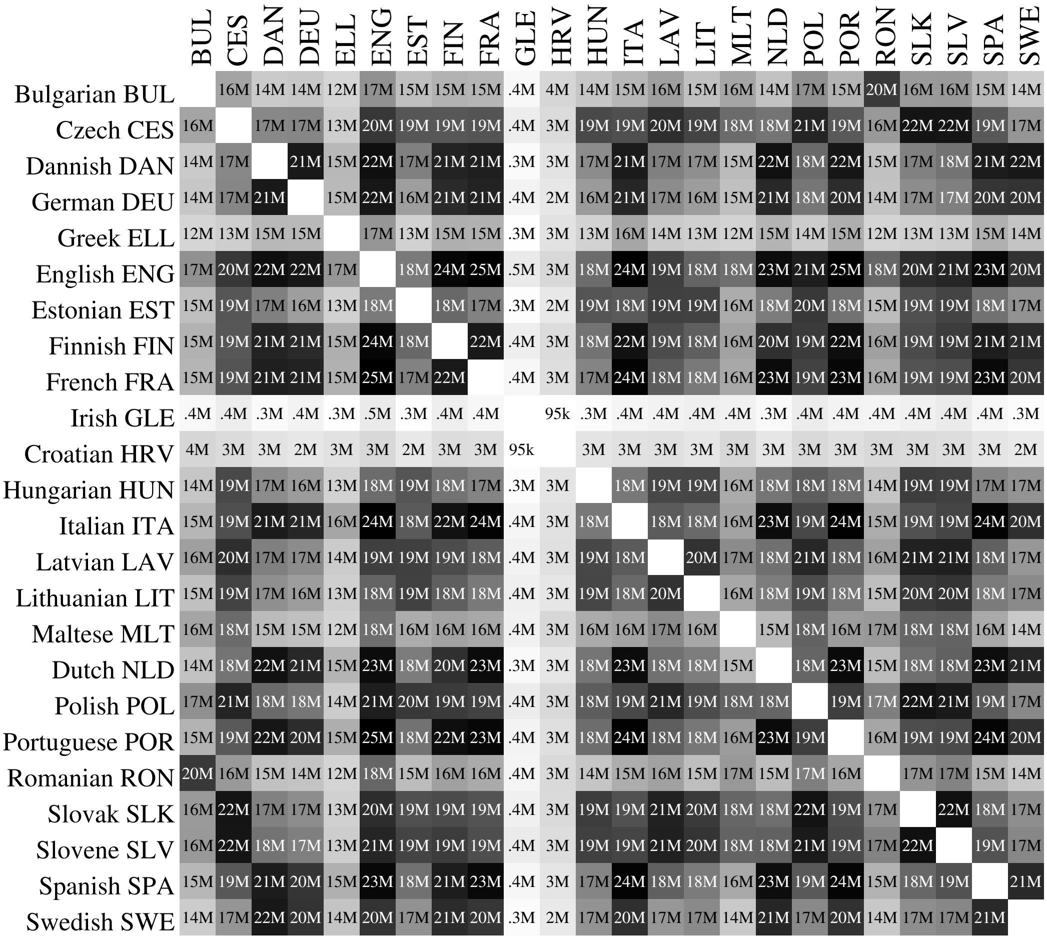
According to the copyright notice[11] on EUR-Lex website: "Except where otherwise stated, reuse of the EUR-Lex data for commercial or non-commercial purposes is authorised provided the source is acknowledged © *European Union, http://eur-lex.europa.eu/, 1998–2015*". This allows us to provide the downloaded data to researchers.[2] Fully processed data (tokenized, PoS-tagged) is not available due to taggers' copyright reasons but available in Sketch Engine.

[8] Norwegian and Icelandic languages are represented in EUR-Lex, but we have omitted them from the final data set due to the negligible number of documents.

[9] http://eur-lex.europa.eu/statistics/eu-law-statistics.html

[10] http://eur-lex.europa.eu/legal-content/EN-ES-FR/TXT/?qid=1445777763012&uri=CELEX:32013R1303&from=EN

[11] http://eur-lex.europa.eu/content/legal-notice/legal-notice.html



	BUL	CES	DAN	DEU	ELL	ENG	EST	FIN	FRA	GLE	HRV	HUN	ITA	LAV	LIT	MLT	NLD	POL	POR	RON	SLK	SLV	SPA	SWE
Bulgarian BUL		16M	14M	14M	12M	17M	15M	15M	15M	.4M	4M	14M	15M	16M	15M	16M	14M	17M	15M	20M	16M	16M	15M	14M
Czech CES	16M		17M	17M	13M	20M	19M	19M	19M	.4M	3M	19M	19M	20M	19M	18M	18M	21M	19M	16M	22M	22M	19M	17M
Dannish DAN	14M	17M		21M	15M	22M	17M	21M	21M	.3M	3M	17M	21M	17M	17M	15M	22M	18M	22M	15M	17M	18M	21M	22M
German DEU	14M	17M	21M		15M	22M	16M	21M	21M	.4M	2M	16M	21M	17M	16M	15M	21M	18M	20M	14M	17M	17M	20M	20M
Greek ELL	12M	13M	15M	15M		17M	13M	15M	15M	.3M	3M	13M	16M	14M	13M	12M	15M	14M	15M	12M	13M	13M	15M	14M
English ENG	17M	20M	22M	22M	17M		18M	24M	25M	.5M	3M	18M	24M	19M	18M	18M	23M	21M	25M	18M	20M	21M	23M	20M
Estonian EST	15M	19M	17M	16M	13M	18M		18M	17M	.3M	2M	19M	18M	19M	19M	16M	18M	20M	18M	15M	19M	19M	18M	17M
Finnish FIN	15M	19M	21M	21M	15M	24M	18M		22M	.4M	3M	18M	22M	19M	18M	16M	20M	19M	22M	16M	19M	19M	21M	21M
French FRA	15M	19M	21M	21M	15M	25M	17M	22M		.4M	3M	17M	24M	18M	18M	16M	23M	19M	23M	16M	19M	19M	23M	20M
Irish GLE	.4M	.4M	.3M	.4M	.3M	.5M	.3M	.4M	.4M		95k	.3M	.4M	.4M	.4M	.4M	.3M	.4M	.4M	.4M	.4M	.4M	.4M	.3M
Croatian HRV	4M	3M	3M	2M	3M	3M	2M	3M	3M	95k		3M	3M	3M	3M	3M	3M	3M	3M	3M	3M	3M	3M	2M
Hungarian HUN	14M	19M	17M	16M	13M	18M	19M	18M	17M	.3M	3M		18M	19M	19M	16M	18M	18M	18M	14M	19M	19M	17M	17M
Italian ITA	15M	19M	21M	21M	16M	24M	18M	22M	24M	.4M	3M	18M		18M	18M	16M	23M	19M	24M	15M	19M	19M	24M	20M
Latvian LAV	16M	20M	17M	17M	14M	19M	19M	19M	18M	.4M	3M	19M	18M		20M	17M	18M	21M	18M	16M	21M	21M	18M	17M
Lithuanian LIT	15M	19M	17M	16M	13M	18M	19M	18M	18M	.4M	3M	19M	18M	20M		16M	18M	19M	18M	15M	20M	20M	18M	17M
Maltese MLT	16M	18M	15M	15M	12M	18M	16M	16M	16M	.4M	3M	16M	16M	17M	16M		15M	18M	16M	17M	18M	18M	16M	14M
Dutch NLD	14M	18M	22M	21M	15M	23M	18M	20M	23M	.3M	3M	18M	23M	18M	18M	15M		18M	23M	15M	18M	18M	23M	21M
Polish POL	17M	21M	18M	18M	14M	21M	20M	19M	19M	.4M	3M	18M	19M	21M	19M	18M	18M		19M	17M	22M	21M	19M	17M
Portuguese POR	15M	19M	22M	20M	15M	25M	18M	22M	23M	.4M	3M	18M	24M	18M	18M	16M	23M	19M		16M	19M	19M	24M	20M
Romanian RON	20M	16M	15M	14M	12M	18M	15M	16M	16M	.4M	3M	14M	15M	16M	15M	17M	15M	17M	16M		17M	17M	15M	14M
Slovak SLK	16M	22M	17M	17M	13M	20M	19M	19M	19M	.4M	3M	19M	19M	21M	20M	18M	18M	22M	19M	17M		22M	18M	17M
Slovene SLV	16M	22M	18M	17M	13M	21M	19M	19M	19M	.4M	3M	19M	19M	21M	20M	18M	18M	21M	19M	17M	22M		19M	17M
Spanish SPA	15M	19M	21M	20M	15M	23M	18M	21M	23M	.4M	3M	17M	24M	18M	18M	16M	23M	19M	24M	15M	18M	19M		21M
Swedish SWE	14M	17M	22M	20M	14M	20M	17M	21M	20M	.3M	2M	17M	20M	17M	17M	14M	21M	17M	20M	14M	17M	17M	21M	

Figure 4: Aligned paragraph counts in EUR-Lex corpus. Millions (M) and thousands (k), darker means larger alignment.

Since EUR-Lex documents contain rich meta data, various aspect can be studied in Sketch Engine. E.g. one can study the trends in keywords and translations in last 60 years, discover language characteristics per EU body, extract domain terminologies using EuroVoc thesaurus etc. We will leave the enumerating of all the possibilities for the reader.

6. Conclusion

We have described a few European multilingual resources and how we made them available in the corpus manager Sketch Engine for lexicographers, linguists and language researchers in general. This allows them to search the full text data using a rich query language which is more suitable for linguistically motivated searches than the full text search engine used on EUR-Lex official web page. Users can also use various statistics derived from the data, e.g. distributional thesaurus, automatic collocations, keyword and terminology candidates, bilingual terminology candidates, parallel collocates and much more.

We have also described a new resource—EUR-Lex corpus—which is to our knowledge the largest resource built from EU data at the moment. Thanks to the permissive data policy of EU we can provide the full data to researchers.[2]

In the future, we plan to download and process EUR-Lex documents also in other formats (PDF, DOCX). This should yield even more parallel data. Another way of getting more parallel data is just to repeat the whole processing once every few months since the EU Publication Office adds new documents to EUR-Lex every day.

7. Acknowledgements

This work has been partly supported by the Grant Agency of CR within the project 15-13277S. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047, LINDAT/CLARIN project LM2010013 and within the Specific University Research.

8. References

Baisa, V., Ulipová, B., and Cukr, M. (2015). Bilingual terminology extraction in sketch engine. In Aleš Horák, Pavel Rychlý, A. R., editor, *Ninth Workshop on Recent Advances in Slavonic Natural Language Processing*, pages 61–67, Brno. Tribun EU.

Carreras, X., Chao, I., Padró, L., and Padró, M. (2004). Freeling: An open-source suite of language analyzers. In *LREC*.

Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102.

Document type	Docs	Author	Docs	EuroVoc	Docs	Year	Docs
Written question	156,744	European Commission	150,545	State aid	18,239	2013	24,978
Regulation	59,758	European Parliament	104,323	European Commission	18,057	2011	24,852
judicial information	36,964	Provisional data	53,230	information transfer	15,778	2012	22,879
Decision	20,400	Council of the EU	31,453	control of State aid	14,096	2010	22,266
Question at Question Time	19,027	Court of Justice	22,397	import	14,074	2007	20,216
Communication	16,384	Court of Justice of the EU	14,637	econ. concentration	12,620	2008	19,238
Consolidated text	16,060	Court of First Instance	12,201	merger control	12,558	2009	18,088
decision w/out addressee	13,718	General Court	9,056	originating product	11,896	2006	17,822
Judgment	13,709	EES Committee	4,524	Italy	11,831	2003	16,587
Proposal for a regulation	8,608	United Kingdom	3,995	Spain	10,882	2005	16,407
Opinion	7,774	EEA Joint Committee	2,880	annul. of EC decis.	10,698	2000	16,248
National exec. measures	7,745	Civil Service Tribunal	2,830	EU Member State	10,562	2001	16,044
Information	7,314	Malta	2,184	Germany	10,274	1996	15,293
Notice	7,306	The Member States	1,978	interpr. of the law	10,030	2004	14,974
Adv. General's Opinion	7,155	Ireland	1,729	EU programme	9,760	1998	14,946
Treaty	5,808	National Courts	1,674	export refund	9,337	1997	14,929
Own-initiative resolution	5,460	Committee of the Regions	1,364	award of contract	9,258	2014	14,868
Report	4,454	European Court of Auditors	1,248	third country	9,210	2002	14,868
Implementing regulation	4,205	The 12 Member States	1,182	trademark law	9,110	1995	14,319
proposal for a decision	4,066	EFTA Surveillance Authority	985	European trademark	8,912	1999	12,667
Info	4,066	European Central Bank	847	environ. protection	8,693	1992	10,768
Directive	3,795	KOSTOPOULOS	807	EU financing	8,212	1993	9,693
Order	3,407	Others	686	import (EU)	8,060	1986	9,265
Own-initiative report	3,054	Gov. representatives	639	EU aid	8,015	1990	9,259
Opinion proposing amend.	3,039	The 6 Member States	622	France	7,980	1985	9,224

Table 3: Example of meta data in English part of EUR-Lex corpus, sorted by document frequency.

Hajlaoui, N., Kolovratnik, D., Väyrynen, J., Steinberger, R., and Varga, D. (2014). Dcep-digital corpus of the european parliament. In *LREC*, pages 3164–3171.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). Hunpos: an open source trigram tagger. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 209–212. Association for Computational Linguistics.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Michelfeit, J., Pomikálek, J., and Suchomel, V. (2014). Text tokenisation using unitok. In *8th Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU*, pages 71–75.

Schmid, H. (1995). Treetaggerl a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43:28.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*.

Steinberger, R., Eisele, A., Klocek, S., Pilos, S., and Schlüter, P. (2013). Dgt-tm: A freely available translation memory in 22 languages. *arXiv preprint arXiv:1309.5226*.

Tiedemann, J. (2009). News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Varga, D., Halácsy, P., Kornai, A., Nagy, V., Németh, L., and Trón, V. (2007). Parallel corpora for medium density languages. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 292:247.