DCEP -Digital Corpus of the European Parliament

Najeh Hajlaoui[1], David Kolovratnik[1], Jaakko Väyrynen[2], Ralf Steinberger[2], Daniel Varga[3]
European Parliament[1], European Commission[2], Budapest University of Technology[3]
DGTRAD Luxembourg[1], JRC Ispra Italy[2], MRC Budapest Hungary[3]
E-mail: {najeh.hajlaoui david.kolovratnik}@ext.europarl.europa.eu[1] {jaakko.vaeyrynen ralf.steinberger}@jrc.ec.europa.eu[2] daniel@mokk.bme.hu[3]

Abstract

We are presenting a new highly multilingual document-aligned parallel corpus called DCEP - Digital Corpus of the European Parliament. It consists of various document types covering a wide range of subject domains. With a total of 1.37 billion words in 23 languages (253 language pairs), gathered in the course of ten years, this is the largest single release of documents by a European Union institution. DCEP contains most of the content of the European Parliament's official Website. It includes different document types produced between 2001 and 2012, excluding only the documents already exist in the Europarl corpus to avoid overlapping. We are presenting the typical acquisition steps of the DCEP corpus: data access, document alignment, sentence splitting, normalisation and tokenisation, and sentence alignment efforts. The sentence-level alignment is still in progress but based on some first experiments; we showed that DCEP is very useful for NLP applications, in particular for Statistical Machine Translation.

Keywords: European Parliament, corpus, European langages

1. Introduction

In 2003, the European Parliament and the Council formulated their insight[1] that public sector information, including raw language data, are useful primary material for digital content products and services, but documents were not initially freely accessible. Since (Koehn, 2005) released his *EuroParl* sentence-aligned data in initially 11 languages and now available in 21 languages[2], such European Union (EU) text material has been widely used to train Statistical Machine Translation (SMT) systems and more. When the European Commission's *Joint Research Centre* (JRC) released the 23-language JRC-Acquis sentence-aligned parallel corpus JRC-Acquis in 2006 (Steinberger, et al., 2006), an SMT system was trained for 462 language pair directions (Koehn, et al., 2009). Several other EU corpora have followed since (Steinberger, et al., 2013).

A limitation of most of these corpora is linked to the administrative text type: while they contain wide-coverage vocabulary – ranging from economy to social issues, science, education, sports, trade and more – their register and text style is rather limited. DCEP – which does not contain the verbatim reports of the EP's plenary sessions already released by Koehn – includes a wider variety of text types. Especially the approximately 12% of press releases should be useful due to their media language.

The corpus is currently aligned at document level and work is on-going to sentence-align it for all language pairs so that data ready to be used to train SMT systems will be ready for distribution as soon as they have been produced. The following sections describe the DCEP collection in detail (Section 2) and they list some of the possible uses of this data (Section 3). We conclude with pointers to forthcoming work.

2. DCEP Collection

The Digital Corpus of the European Parliament (DCEP) contains most of the content of the European Parliament's official Website[3]. It includes the following different document types produced between 2001 and 2012:

- AGENDA: Agenda of the plenary session meetings;
- COMPARL: Draft Agenda of the part-session;
- IM-PRESS and PRESS: General texts and articles on parliamentary news seen from a national angle, specific to one or several Member States, presentation of events in the EP;
- IMP-CONTRIB: Various press documents including technical announcements, events (hearings, workshops) produced by the Parliamentary Committees;
- MOTION: Motions for resolutions put to the vote in plenary;
- PV: Minutes of plenary sittings;
- REPORT: Reports of the parliamentary committees;
- RULES-EP: The Rules of Procedure of the EP laying down the rules for the internal operation and organisation of EP;
- TA (Adopted Texts): The motions for resolutions and reports tabled by Members and by the parliamentary committees are put to the vote in plenary, with or without a debate. After the vote, the final texts as adopted are published and forwarded to the authorities concerned;
- WQ (Written Question), WQA (Written Question Answer), OQ (Oral Question) and QT

[1] Directive 2003/98/EC of the European Parliament and of the Council on the re-use of public sector information: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003L0098:EN:NOT

[2] See http://www.statmt.org/europarl/

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

(Questions for Question Time).

As explained in (Koehn, 2005), the acquisition of a parallel corpus typically takes the same steps: data access, document alignment, sentence splitting, normalisation and tokenisation, and sentence alignment.

2.1 Data access

Contrary to the crawling method used to build the Europarl corpus (Koehn, 2005), the DCEP corpus is downloaded directly from the in-house database of the European Parliament. The motivation behind the DCEP collection is to offer the NLP community a unique multilingual corpus different in terms of size and in terms of content variety from the previous published corpora (Steinberger, et al., 2013).

The CRE "Compte Rendu in Extenso" documents are not included in the DCEP corpus to avoid overlapping with the Europarl corpus. CRE are the verbatim reports of the speeches made in the European Parliament's plenary.

	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	GA
CS	14 341										
DA	14 626	19 961									
DE	14 825	19 910	102 581								
EL	14 804	20 114	101 559	101 737							
EN	15 204	20 597	104 260	107 760	109 090						
ES	14 823	20 191	102 833	103 017	101 868	107 079					
ET	14 213	19 677	19 632	19 454	19 748	20 010	19 793				
FI	14 788	19 499	101 987	102 554	101 004	102 830	102 256	19 065			
FR	14 891	20 048	102 775	103 688	102 506	109 845	103 814	19 613	102 421		
GA	12	12	13	13	13	14	13	12	13	13	
HU	14 557	19 531	19 802	20 067	20 018	20 603	20 141	19 521	19 712	20 166	12
IT	14 780	20 158	102 803	102 999	101 954	109 411	103 222	19 746	102 195	103 964	13
LT	14 457	19 737	20 164	20 142	20 322	20 912	20 424	19 786	19 708	20 318	12
LV	14 413	19 748	19 766	19 626	19 882	20 179	19 964	19 857	19 190	19 769	12
MT	14 033	17 030	17 506	17 485	17 660	18 213	17 672	17 176	17 229	17 610	12
NL	14 701	20 026	102 767	102 901	101 759	107 115	103 025	19 687	102 081	103 439	13
PL	14 387	19 612	21 068	21 090	21 227	22 630	21 302	19 610	20 779	21 270	12
PT	14 677	19 767	102 413	102 686	101 524	105 566	102 858	19 418	102 278	103 181	13
RO	14 562	14 897	16 035	15 954	16 221	17 526	16 286	14 851	15 380	16 244	12
SK	14 431	19 597	19 940	20 022	20 142	20 946	20 181	19 605	19 873	20 096	12
SL	14 319	19 461	19 419	19 591	19 628	19 846	19 663	19 440	19 332	19 653	12
SV	14 670	20 086	102 738	102 709	101 673	103 831	102 937	19 791	102 183	102 836	13
	HU	**IT**	**LT**	**LV**	**MT**	**NL**	**PL**	**PT**	**RO**	**SK**	**SL**
IT	20 058										
LT	19 691	20 356									
LV	19 657	19 895	19 931								
MT	17 007	17 652	17 134	17 244							
NL	19 904	103 031	20 197	19 809	17 575						
PL	19 621	21 251	19 723	19 733	17 025	21 157					
PT	19 855	102 728	19 967	19 515	17 403	102 574	20 989				
RO	14 857	16 293	15 016	15 068	14 462	16 203	14 846	15 812			
SK	19 612	20 128	19 701	19 692	17 123	20 030	19 623	19 868	14 768		
SL	19 552	19 599	19 565	19 544	16 964	19 548	19 472	19 377	14 713	19 503	
SV	19 919	102 925	20 285	19 910	17 606	102 876	21 198	102 523	16 139	20 071	19 593

Table 1: Number of documents per language pair.

2.2 Document Alignment

DCEP contains some original texts in SGML and others in XML format. Both are structured by language and by document type. It contains an index file showing links between linguistic versions of documents. Based on this index, we created a bilingual corpus and it can be used also to create multilingual corpora. This index allowed us to present the following statistical information: it contains a space-separated list of file names[4] of corresponding linguistic versions of documents. For instance, if there is only one file name, it means that the document is available only in one language. Because it happens that more than one linguistic version for the same document (and for the same language) exists, we excluded them for the case of multilingual corpora but we included them to

[4] Example of file name: 16338845__IM-PRESS__20050826-IPR-01421__EN

build a monolingual corpus or to present statistical details.

	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	GA
BG											
CS	32 565										
DA	32 380	41 835									
DE	33 945	43 365	74 238								
EL	35 109	45 123	77 164	79 230							
EN	35 325	45 333	77 522	80 929	84 352						
ES	37 144	48 039	82 198	84 370	86 941	88 597					
ET	29 790	38 280	38 395	39 817	41 580	41 719	44 436				
FI	29 579	37 704	64 555	66 869	69 395	69 555	74 442	34 039			
FR	36 399	47 053	80 807	83 050	85 607	90 312	**91 016**	43 462	73 027		
GA	1 123	1 044	1 086	1 120	1 151	1 160	1 212	991	980	1 183	
HU	31 968	40 698	40 946	42 831	44 184	44 482	47 078	37 345	36 969	46 159	1 053
IT	35 564	46 089	78 496	80 688	83 432	85 769	88 721	42 415	70 960	87 480	1 157
LT	30 990	39 669	39 978	41 273	43 019	43 284	46 074	36 130	35 570	45 270	1 022
LV	31 196	39 769	39 840	41 350	43 114	43 358	46 057	36 340	35 592	45 050	1 028
MT	31 734	36 297	36 227	37 876	39 222	39 869	41 928	33 170	32 628	41 054	1 065
NL	34 776	44 729	77 544	79 554	82 182	83 754	87 532	41 327	69 687	86 089	1 173
PL	32 660	41 865	42 066	43 925	45 339	45 946	48 393	38 333	38 120	47 396	1 061
PT	35 564	45 487	78 475	80 598	83 359	84 327	88 373	41 990	70 989	87 080	1 167
RO	34 766	33 048	33 042	34 530	35 833	36 300	37 945	30 347	29 936	37 147	1 137
SK	32 374	41 507	41 771	43 197	44 829	45 211	47 967	38 064	37 564	47 011	1 051
SL	32 201	41 418	41 386	43 007	44 722	44 838	47 639	37 945	37 359	46 655	1 045
SV	32 812	42 381	72 724	74 829	77 572	77 957	82 834	38 816	65 157	81 103	1 089
	HU	**IT**	**LT**	**LV**	**MT**	**NL**	**PL**	**PT**	**RO**	**SK**	**SL**
IT	45 009										
LT	38 628	43 964									
LV	38 814	44 074	37 768								
MT	35 473	40 289	34 382	34 721							
NL	43 870	83 799	42 917	42 802	39 018						
PL	41 080	46 338	39 767	39 921	36 470	45 088					
PT	44 495	84 966	43 439	43 578	39 683	83 553	45 809				
RO	32 314	36 330	31 501	31 763	32 165	35 525	33 124	36 225			
SK	40 610	45 799	39 539	39 650	36 203	44 770	41 709	45 312	32 787		
SL	40 470	45 555	39 200	39 481	35 967	44 339	41 441	45 095	32 703	41 239	
SV	41 362	79 166	40 189	40 381	36 866	77 983	42 628	78 929	33 527	42 174	42 014

Table 2: Average number of words per language pair (in thousands)

Table 1, Table 2 and Table 3 respect ively present, number of documents, average of the number of words, and average of the number of unique words per language[5] pair.

The French-Spanish language pair has the most words and the following pairs have at most 10% less: Greek, English, Spanish and French paired to Italian, Dutch and Portuguese; French and Spanish paired with German, Greek and English; also Spanish paired with Danish, French and Swedish; and finally English-Greek, Dutch-Italian, Portuguese-Italian and Portuguese-Dutch.

2.3 Sentence splitting and tokenisation

In order to split documents into sentences, we followed two steps: the first consists of replacing the structural mark-up by a new line rather than deleting it. Table 4 shows why respecting the document structure is important for segmentation. For each document type such tags were selected manually. Besides this, again just for selected document types, line breaks are promoted from within a tag in order to act as a segment separator. Line breaks from the document are preserved as well. The second step consists of using the Moses script to separate sentences if they still appear on one line. The script was modified so that it never merges any segment spread across more lines.

General statistics on the documents, words and sentences are shown in Table 5: for each language. There are more than 100,000 documents for the languages of the member states prior to 1995 (DA, DE, EL, EN, ES, FI, FR, IT, NL, PT, and SV). There are about 20,000 documents for the languages of member states that joined in 2004 or after (BG, CS, ET, HU, LT, LV, MT, PL, RO, SK, and SL). The Turkish language (TR) has very few documents compared to the others. GA (Irish) has more than one million words, whereas there is basically no material for TR. The differences in language productivity are measured by the

[5] We are using the iso-639-1language code (http://www.iso.org/iso/language_codes).

	BG	CS	DA	DE	EL	EN	ES	ET	FI	FR	GA	HU	IT	LT	LV	MT	NL	PL	PT	RO	SK	SL
BG																						
CS	138																					
DA	172	177																				
DE	180	184	343																			
EL	171	172	314	323																		
EN	132	125	245	258	228																	
ES	137	132	256	264	236	166																
ET	208	228	260	267	255	208	215															
FI	235	252	448	459	428	359	369	335														
FR	127	120	239	247	219	169	159	203	352													
GA	5	3	4	4	5	4	4	5	5	3												
HU	132	137	170	178	165	118	125	220	245	113	2											
IT	146	142	268	277	248	180	188	225	383	171	4	134										
LT	195	210	245	251	239	192	200	294	318	188	6	204	208									
LV	185	196	229	237	224	177	185	280	304	173	6	189	194	264								
MT	155	143	176	183	173	129	136	221	245	124	4	136	145	205	194							
NL	174	178	334	342	313	246	255	261	447	238	4	172	267	246	231	178						
PL	183	192	228	236	223	176	182	276	304	170	6	186	192	258	245	189	229					
PT	135	129	251	260	232	162	172	212	366	155	3	122	185	195	181	133	250	179				
RO	162	126	160	168	160	120	125	197	224	115	5	120	134	184	174	143	163	171	123			
SK	142	147	181	187	175	128	135	231	255	123	3	140	145	214	200	147	182	195	132	130		
SL	176	185	217	224	212	165	172	269	293	160	6	178	182	251	237	181	219	233	169	164	188	
SV	162	164	313	322	293	223	234	248	429	216	3	157	246	231	216	164	312	215	230	150	167	205

Table 3: Average number of unique words per language pair (in thousands)

Before splitting
<Infopress language="EN" xmlns:ns1="SipadeType" xmlns:xhtml="http://www.w3.org/1999/xhtml"><Title>EU should cooperate more with US in Mediterranean region</Title><Topic>Development and cooperation</Topic><PublicationDate>2005-09-07 - 18:26</PublicationDate><Photography href="20050822PHT01307" title=" " alt=" " ext="jpg" width="697" height="501"> </Photography>
After splitting: 3 sentences (Title, Topic and PublicationDate**)**
EU should cooperate more with US in Mediterranean region Development and cooperation 2005-09-07 - 18 : 26

Table 4: Example of sentence splitting

Standardized Type/Token Ratio (STTR[6]), which enables comparison of corpora with different lengths. FI and ET are morphologically generative languages and have the highest values of STTR. The lower values are with ES and GA.

The best-represented language in terms of number of words is English. Comparatively, French and Spanish miss less than 10%. On the other hand, each language has at least 30 % of the English number of words, only Bulgarian and Estonian are below 35%.

2.4 DCEP Word Distributions

The numbers of words in documents for each language are summarized in Table 6, which shows selected percentiles[7], the mean and the standard deviation (Std). A majority of the documents have less than 5,000 words, but there are some much longer documents. The more recent members of the EU have proportionally longer documents than the older member states. Compared to the other languages, GA does not have very short documents at all, whereas there are only very short documents in TR.

For the purpose of statistical machine translation, sentences are the main translation units. The number of sentences in documents is relevant for efficient sentence alignment, and the total number of sentences and sentence lengths are relevant for word alignment and resource management. Table 7 shows statistics on the number of sentences in each language without cross-lingual alignments. The conclusions of the analysis are similar to those of Table 6. Table 8 shows statistics on the number of words in sentences. Half of the sentences are very short with at most 3-5 words. There are some very long sentences, but nearly all are below the typical threshold of 80 or 100 words.

2.5 Sentence Alignment

We are creating sentence alignments for all documents and all possible language pairs of the DCEP. A part of this work is already completed for some language pairs such as EN/FR. This meant a very large number of alignments, so we had to choose a fast alignment algorithm. We used the HunAlign sentence aligner (Varga, et al., 2007), a common choice among creators of large multilingual parallel corpora (Tiedemann, 2009) (Waldenfels, 2011) (Rosen, et al., 2012).

[6] STTR = TTR computed after each block of n words, here n = 1000, then we took the average of all blocks TTR. Tokens were strings separated by whitespaces, while types were unique strings of those.

[7] 0th percentile gives the length of the shortest document and the 100th the length of the longest document.

Language	#documents	# sentences	# words	# unique words	STTR
BG	15,881	3,189,893	35,265,634	533,756	47.22%
CS	21,211	4,457,637	42,732,357	707,055	54.35%
DA	105,138	6,709,190	74,034,195	1,335,980	47.50%
DE	109,644	6,545,600	79,956,002	1,314,460	47.99%
EL	110,931	6,778,311	86,851,326	1,108,140	48.28%
EN	162,608	7,650,837	**103,458,996**	1,049,826	44.63%
ES	108,691	6,590,119	95,457,198	911,105	41.95%
ET	20,538	4,072,770	35,319,468	947,169	58.24%
FI	104,513	6,348,983	58,274,608	1,802,139	61.55%
FR	115,881	6,914,801	98,630,448	1,004,068	44.96%
GA	14	123,968	1,222,234	11,219	41.68%
HU	21,543	4,196,424	41,277,563	971,455	53.52%
IT	111,195	6,737,167	89,099,402	1,010,644	48.10%
LT	21,589	4,265,335	38,703,299	733,480	56.97%
LV	20,705	4,212,867	38,587,221	713,506	55.43%
MT	18,819	3,804,307	36,593,231	761,320	54.73%
NL	108,402	6,527,499	85,787,172	1,187,851	42.84%
PL	23,466	4,152,915	43,647,099	746,864	54.80%
PT	107,175	6,442,722	88,065,967	953,049	45.34%
RO	17,777	3,083,763	36,270,771	534,468	48.99%
SK	21,841	4,281,697	42,536,235	713,273	54.64%
SL	20,633	4,193,239	41,844,125	668,778	53.64%
SV	104,665	6,548,318	74,501,242	1,255,700	47.90%
TR	6	24	56	17	N/A

Table 5: The number of documents, sentences, words (tokens), unique words (types), and STTR for each language.

	Percentiles for the number of words in documents								
Language	0th	10th	25th	50th	75th	90th	100th	Mean	Std
BG	7	146	275	634	2,071	5,628	183,614	2,220.6	5,493.6
CS	3	125	247	619	1,927	5,073	157,440	2,014.6	4,850.4
DA	0	65	120	199	344	1,261	134,803	704.2	2,560.9
DE	3	72	126	208	362	1,254	247,799	729.2	2,811.6
EL	3	80	140	229	399	1,328	192,797	782.9	2,889.5
EN	3	72	135	233	396	877	178,840	636.2	2,424.5
ES	3	87	152	248	432	1,504	153,137	878.2	3,224.4
ET	3	105	211	531	1,649	4,330	134,779	1,719.7	4,185.5
FI	3	53	92	151	266	987	195,439	557.6	2,183.1
FR	3	83	148	247	441	1,454	171,177	851.1	3,100.1
GA	150	578	88,143	98,091	113,064	114,857	114,993	87,302.4	38,185.7
HU	3	120	228	565	1,769	4,793	236,428	1,916.1	4,989.5
IT	3	78	138	226	392	1,346	181,047	801.3	3,019.1
LT	3	114	215	542	1,717	4,580	125,500	1,792.7	4,306.2
LV	3	118	234	577	1,809	4,665	148,838	1,863.7	4,550.4
MT	0	109	228	576	1,797	4,935	178,826	1,944.5	4,889.9
NL	3	78	138	227	394	1,373	144,227	791.4	2,908.0
PL	3	118	213	519	1,658	4,735	237,051	1,860.0	4,969.9
PT	3	81	141	232	400	1,448	194,605	821.7	3,021.6
RO	7	134	243	557	1,751	5,189	180,895	2,040.3	5,335.4
SK	3	121	227	579	1,818	4,910	133,796	1,947.5	4,748.7
SL	3	125	251	636	1,949	5,051	173,137	2,028.0	4,921.8
SV	3	67	118	195	340	1,258	166,282	711.8	2,669.6
TR	9	9	9	9	10	10	10	9.3	0.7

Table 6: Bowley's seven-number summary, the mean and standard deviation for the number of words in documents for each language.

Language	Percentiles for the number of sentences in documents							Mean	Std
	0th	10th	25th	50th	75th	90th	100th		
BG	1	12	24	61	168	412	11,115	200.9	565.3
CS	1	12	27	62	164	425	26,373	210.2	595.0
DA	0	7	10	14	23	104	26,660	63.8	315.8
DE	1	7	10	13	21	94	25,872	59.7	313.7
EL	1	7	9	13	22	99	26,750	61.1	295.5
EN	1	6	9	13	20	63	26,460	47.1	243.3
ES	1	7	9	13	22	97	26,415	60.6	297.8
ET	1	12	25	58	154	406	26,253	198.3	578.5
FI	1	7	10	14	23	98	26,243	60.7	298.1
FR	1	7	9	13	24	94	35,246	59.7	311.5
GA	11	19	9,809	10,261	11,231	11,396	11,481	8,854.9	3,917.7
HU	1	11	23	56	153	405	26,212	194.8	595.9
IT	1	7	9	13	22	95	26,264	60.6	316.4
LT	1	12	24	58	163	411	27,045	197.6	573.7
LV	1	12	25	60	163	415	26,324	203.5	594.9
MT	0	10	22	57	160	439	26,381	202.2	586.9
NL	1	7	10	14	22	96	26,373	60.2	301.0
PL	1	10	18	48	136	351	26,314	177.0	558.5
PT	1	7	9	13	21	97	26,310	60.1	296.8
RO	1	10	18	46	135	326	12,579	173.5	565.0
SK	1	11	22	56	157	400	26,399	196.0	577.3
SL	1	12	24	59	163	413	26,223	203.2	587.3
SV	1	7	9	13	21	105	26,300	62.6	298.0
TR	4	4	4	4	4	4	4	4.0	N/A

Table 7: Bowley's seven-number summary, the mean and standard deviation for the number of sentences in documents for each language.

In employing HunAlign for our corpus, we followed the approach of the JRC-Acquis corpus (Steinberger, et al., 2006). For a single language pair, this workflow consists of an initial alignment of all document pairs, a sampling of the identified sentence pairs, a dictionary-building phase based on the sentence pairs, and finally a second alignment that considers the automatic dictionary when calculating sentence similarity. We note that for calculating similarity, HunAlign employs heuristics that compare the sets of number tokens found in the source and target sentences, an especially relevant clue when aligning legal text such as DCEP, where a significant percentage of the sentences contain number tokens.

We altered the JRC-Acquis workflow slightly, because the DCEP contains some very long documents that could have slowed down the alignment process. For documents with more than 20,000 sentences we employed partialAlign, a companion tool for HunAlign that splits a document pair into smaller document pairs compatible with the alignment. This shrinks HunAlign's running time and memory consumption significantly, without affecting precision (Varga, 2012).

For the JRC-Acquis corpus the authors provided alignments both by the Vanilla (Gale, et al., 1991) and the hunalign aligner implementations. A manual evaluation of a small sample of this dataset (Kaalep, et al., 2007) found that HunAlign significantly outperforms Vanilla in precision, so we omitted the Vanilla alignments for DCEP.

3. What is DCEP useful for

DCEP is a multilingual corpus including documents in all official EU languages and it can be used for various language processing and research purposes such as:

- Machine Translation, mainly Statistical Machine Translation (SMT);
- Creation of monolingual or multilingual corpora;
- Translation studies, annotation projection for co-reference resolution, discourse analysis, comparative language studies;
- Improvement of sentence or word alignment algorithms;
- Cross-lingual information retrieval.

Table 9 shows as first experiments on using DCEP to train SMT have shown that, even for the well-resourced language pair English-French, the quality goes up significantly when adding DCEP to EuroParl for a DCEP test set (without overlap with the training set): BLEU jumps from 27.9 to 39.3 and METEOR from 46.1 to 54.6; The Translation Error Rate TER drops from 56.7 to 47.5. These scores are still increasing for a shared test set (1000 from each corpus). The ACT "Accuracy of Connectives Translation" (Hajlaoui, et al., 2013) scores show also that discourse connectives are better translated with the (DCEP+Europarl) system.

Language	Percentiles for the number of words in sentences							Mean	Std
	0th	10th	25th	50th	75th	90th	100th		
BG	1	1	1	3	14	32	4,267	11.1	21.7
CS	1	1	1	3	12	28	3,312	9.6	17.7
DA	0	1	1	4	16	31	4,375	11.0	18.5
DE	1	1	1	4	18	34	5,358	12.2	20.4
EL	1	1	1	4	19	37	4,029	12.8	20.8
EN	1	1	1	5	21	37	9,522	13.5	22.0
ES	1	1	1	5	21	42	9,682	14.5	25.4
ET	1	1	1	3	11	25	5,474	8.7	17.0
FI	1	1	1	4	13	25	7,183	9.2	16.5
FR	1	1	1	5	21	40	10,669	14.3	24.1
GA	1	1	1	1	14	32	170	9.9	18.6
HU	1	1	1	3	12	28	4,866	9.8	18.8
IT	1	1	1	4	20	37	6,533	13.2	21.8
LT	1	1	1	3	11	26	1,864	9.1	15.8
LV	1	1	1	3	11	26	8,653	9.2	20.0
MT	0	1	1	3	12	28	8,715	9.6	22.4
NL	1	1	1	4	19	37	7,565	13.1	23.3
PL	1	1	1	3	14	30	2,898	10.5	19.5
PT	1	1	1	4	20	39	9,152	13.7	24.4
RO	1	1	1	3	15	35	4,239	11.8	22.9
SK	1	1	1	3	13	28	6,709	9.9	19.4
SL	1	1	1	3	13	29	4,287	10.0	18.9
SV	1	1	1	4	18	31	7,388	11.4	18.2
TR	1	1	1	3	4	4	4	2.3	1.8

Table 8: Bowley's seven-number summary, the mean and standard deviation for the number of words in sentences for each language.

SMT systems	Training set (Nb. sent)	Tuning set: NC2008 (Nb. sent)	BLEU	METEOR	TER	Length	ACT	
							ACTa	ACTa5+6
DCEP TEST SET: 1000 sentences								
Baseline (Europarl)	1964110	2051	27.9	46.1	56.7	86.3	58.3	84
System (Europarl+DCEP)	4514755	2051	39.3	54.6	47.5	85.1	58.3	84
(EUROPARL+DCEP) TEST SET: 2000 sentences								
Baseline (Europarl)	1964110	2051	32.1	50.1	54.6	97.7	56.9	72.7
System (Europarl+DCEP)	4514755	2051	33.8	51.2	52.4	95.4	57.3	73.2
EUROPARL TEST SET: 1000 sentences								
Baseline (Europarl)	1964110	2051	32.8	51.4	54	101.6	56.6	71.1
System (Europarl+DCEP)	4514755	2051	31.8	50	54.1	98.8	57.1	71.7

Table 9: EuroParl-based SMT baseline vs (EuroParl+DCEP)-based SMT system: Metric scores for all English-French systems: jBLEU V0.1.1 (an exact reimplementation of NIST's mteval-v13.pl without tokenization); Meteor V1.4 en on rank task with all default modules not ignoring punctuation; Translation Error Rate (TER) V0.8.0; Hypothesis length over reference length in percent; ACT (V1.7) scores to assess the discourse connectives translations.

SMT systems are implemented using the Moses decoder (Koehn, et al., 2007) with the phrase-based factored translation models (Koehn, et al., 2007). The language models for French were 3- gram ones over EuroParl v7 (Koehn, 2005) for the Baseline system and over a concatenation of it with the DCEP corpus for the system using the IRSTLM toolkit (Federico, et al., 2008). Minimum Error Rate Training (MERT) (Och, 2003) is used to optimize the systems.

4. Conclusion

We presented a new highly multilingual parallel corpus called DCEP. It is four times bigger than the Europarl corpus and larger in terms of variety (thirteen different document types) and number of languages (23 languages). DCEP thus constitutes the largest release of documents by a European Union institution.

Based on some experiments, we showed that DCEP is very useful for NLP applications, in particular for Statistical Machine Translation.

5. Bibliography

Clark, J., Dyer, C., Lavie, A., & Smith, N. (2011). Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. *In Proceedings of ACL-HLT 2011 (46th Annual Meeting of the ACL: Human Language Technologies.* Portland, OR.

Federico, M., Bertoldi, N., & Cettolo, M. (2008). IRSTLM: an open source toolkit for handling large scale language models. *In Proceedings of Interspeech.* Brisbane, Australia.

Gale, W. A., & Church, K. W. (1991). A program for aligning sentences in bilingual corpora. *in 'Meeting of the Association for Computational Linguistics'*, (pp. 177-184).

Hajlaoui, N., & Popescu-Belis, A. (2013). Assessing the accuracy of discourse connective translations: Validation of an automatic metric. *In Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics.* Samos, Greece.

Kaalep, H.-J., & Veskis, K. (2007). Comparing parallel corpora and evaluating their quality. *in Proceedings of MT Summit XI*, (pp. 275-279).

Koehn, P. (2005). A Parallel Corpus for Statistical Machine Translation. *Machine Translation Summit.*

Koehn, P., & Hieu, H. (2007). Factored Translation Models. *In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning (CONLL)*, (pp. 868-876). Prague, Czech Republic.

Koehn, P., Birch, A., & Steinberger, R. (2009). 462 Machine Translation Systems for Europe. *In: Laurie Gerber, Pierre Isabelle, Roland Kuhn, Nick Bemish, Mike Dillinger, Marie-Josée Goulet (eds.): Proceedings of the Twelfth Machine Translation Summit*, (pp. pages 65-72). Ottawa, Canada.

Koehn, P., Hieu, H., Birch, A., Chris, C.-B., Marcello, F., Bertoldi, N., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. *In Proceedings of 45th Annual Meeting of the Association for Computational Linguistics (ACL), Demonstration Session*, (pp. 177-180). Prague, Czech Republic.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. *In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL),* (pp. 160-167). Sapporo, Japan.

Rosen, A., & Vavřín, M. (2012). Building a multilingual parallel corpus for human users, in N. C. C. Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk & S. Piperidis, eds,. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)' European Language Resources Association (ELRA).* Istanbul, Turkey.

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., et al. (2013). An overview of the European Union's highly multilingual parallel corpora. *Journal Language Resources and Evaluation.*

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., et al. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *Proceedings of the 5th International Conference on Language Resources.*

Tiedemann, J. (2009). News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. *in N. Nicolov, G. Angelova & R. Mitkov, eds, 'Recent Advances in Natural Language Processing, Vol. 309 of Current Issues in Linguistic Theory, John Benjamins, Amsterdam & Philadelphia*, 227-248.

Varga, D. (2012). *Natural Language Processing of Large Parallel Corpora.* Budapest, Hungary: PhD Thesis. Eötvös Loránd University.

Varga, D., Péter, H., András, K., Viktor, N., László, N., & Viktor, T. (2007). Parallel corpora for medium density languages. *In Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005, Nicolov, Nicolas, Kalina Bontcheva, Galia Angelova and Ruslan Mitkov (eds.)*, (pp. 247-258).

Waldenfels, R. V. (2011). Recent Developments in Parasol: Breadth for Depth and Xslt Based Web Concordancing with Cwb. *in 'Proceedings of Slovko 2011*, (pp. 156-162). Modra, Slovakia.