The public translation memories and corpora of DG Translation of the European Commission

Spyridon Pilos
Head of language applications sector
DG Translation, Informatics unit

Aristotle University of Thessaloniki, 17 January 2014
Symposium: Corpus-Based Translation Studies with emphasis on parallel literary corpora

Directorate-General for Translation (DGT)

***EU Official languages**: 24*
(with Croatian added on 1 July 2013)

***EC procedural languages**: 3 (EN, FR, DE)*

Staff: 2300 of which 1500 translators and 800 support and management*

***Where**: in Brussels, Luxembourg and in local (field) offices in Member States*

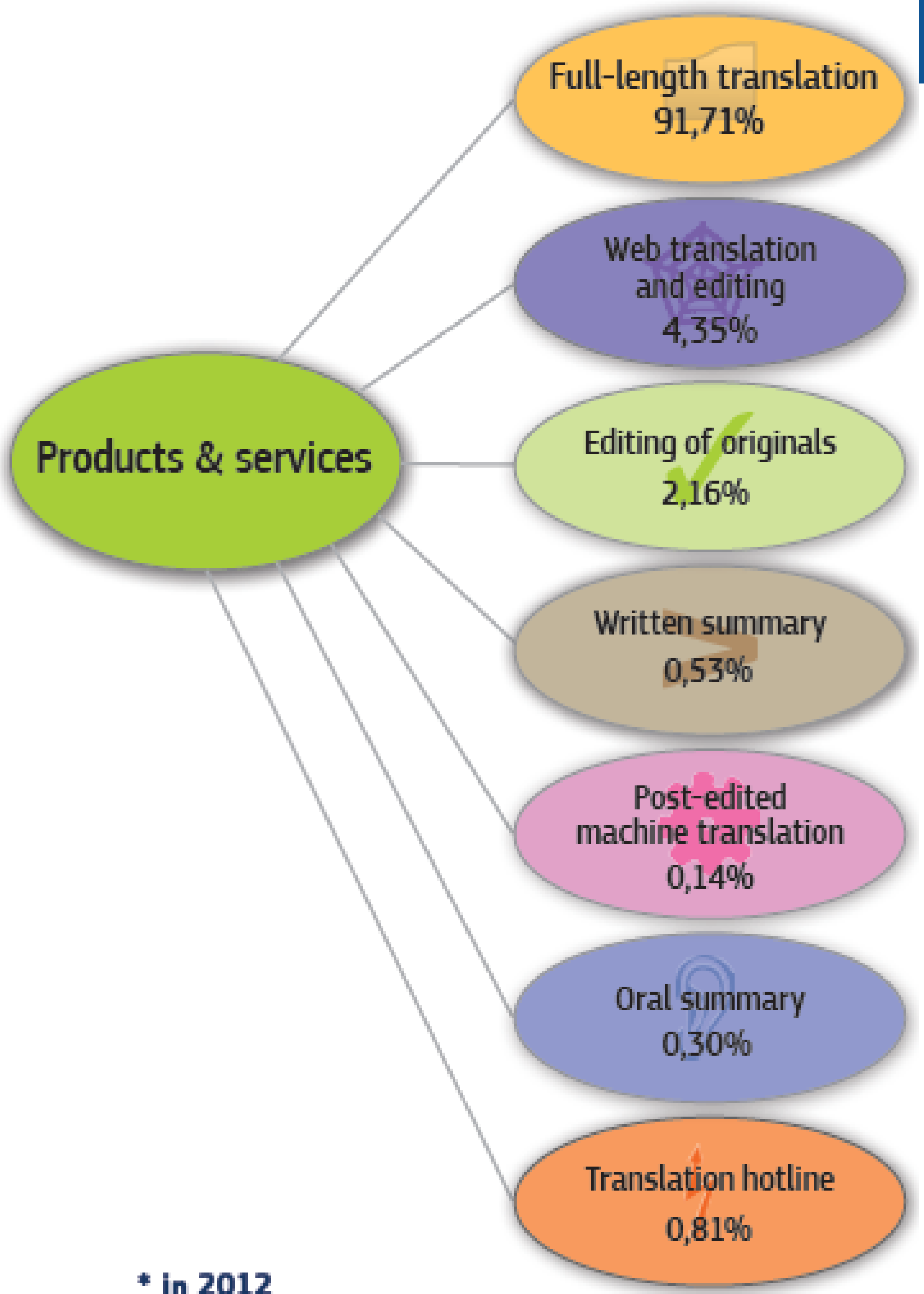
Production: 1 760 615 translated pages*

** 2012 figures*

Target language in %		Pages translated in total
English	14,92%	262 766
French	8,25%	145 303
German	6,47%	113 935
Italian	4,40%	77 415
Spanish	4,26%	75 048
Dutch	4,05%	71 260
Portuguese	3,91%	68 865
Greek	3,81%	67 061
Polish	3,77%	66 384
Slovenian	3,57%	62 831
Swedish	3,57%	62 784
Romanian	3,57%	62 782
Hungarian	3,55%	62 541
Finnish	3,52%	61 905
Slovak	3,52%	61 889
Czech	3,51%	61 855
Bulgarian	3,51%	61 720
Danish	3,45%	60 787
Lithuanian	3,42%	60 262
Latvian	3,41%	59 992
Estonian	3,41%	59 950
Maltese	3,37%	59 388
Irish	0,41%	7 206
Other	0,38%	6 680
	100%	

Pages translated in total: 1 760 615

Distribution across products & services



* in 2012

DGT Translation memories

- **Euramis** (European advanced multilingual information system)
- DGT managing the TMs for all EU institutions
- Currently around 800 million translation units (segments/"sentences") in 24 languages
- Annual growth rate: more 20%

Main source also for building the new statistical machine translation service MT@EC (operational since July 2013).

DGT sharing multilingual data

- *Datasets based on public content*
- *"Acquis communautaire": the* ***biggest parallel corpus*** *in existence (size and number of languages).*
- *Official legal documents already published in the Official Journal(OJ) of the EU and available online (EurLeX)*
- *Covering 23 EU languages (not HR yet)*
- *Documents linguistically and legally checked during a multi-step revision process by DGT, Commission legal services and the EU Publications Office.*

DGT-Acquis

- *multilingual paragraph-aligned parallel corpora*

(i.e. full text documents with added meta-information on which paragraphs are aligned with which others in the other languages).

- *All 23 official EU languages,*
- *documents from the L and C series of OJ*
- *Years covered: 2004 to 2011.*

DGT-Acquis data

- *Processing is four steps (increasing granularity)*
- *The result of each step is a corpus packaged as a self-contained "muset" file:*

  (1) original data,

  (2) file level in Formex4 format,

  (3) file level in plain text and

  (4) paragraph level.
- *"musets" are independent but linked to each other (to find the source document of any given text segment).*

DGT-Acquis - some figures

Title	Granularity	Format	Structure	Zipped	Statistics	Comments
Original data	original	formex4	tree	81 GB	3,901,048 files	original filenames; with TIFF files
File level in Formex4	file	formex4	tree	9 GB	3,537,876 files	standardised filenames; without TIFF files
File level in plain text format	file	text	tree	5 GB	3,537,872 files	XML marking removed
Paragraph level in column-file format	paragraph	column-file	table	3 GB	4,900,254 segments	one table

DGT-TM

- *multilingual **sentence**-aligned translation memory*
- *all 23 official EU languages*
- *documents from the L series of OJ*
- *years covered: 1972 to 2012.*
- *Extracted from Euramis.*
- *Since 2011 automatic transfer (daily) from publications office to Euramis where data are automatically segmented and aligned.*

DGT-TM data

- *Format: TMX (Translation Memory eXchange). Text-encoding: UTF-16 Little Endian.*
- *Source language not identified*
- *You can produce parallel sentence collections for any of the possible language pairs:*

  1st release: 2007 (ref 1972-2006)

  2nd release: 2011 (ref 2007-2010)

  3rd release: 2012 (ref 2011)

  4th release: 2013 (ref 2012)
- *Annual releases planned.*

DGT-TM – comments on content

- the sequence in the extracted files not necessarily the same as in the underlying docs.
- redundancancies of text segments like "*Article 1*".
- Underying docs identified by the doc number (Numdoc) of the original legislative doc in Eur-Lex (although modified during pre-processing)

DGT-TM – some figures

All languages	Translation units
release 2007	19.071.485
release 2011	37.963.629
release 2012	6.226.855
release 2013	10.154.534
TOTAL	73.416.503

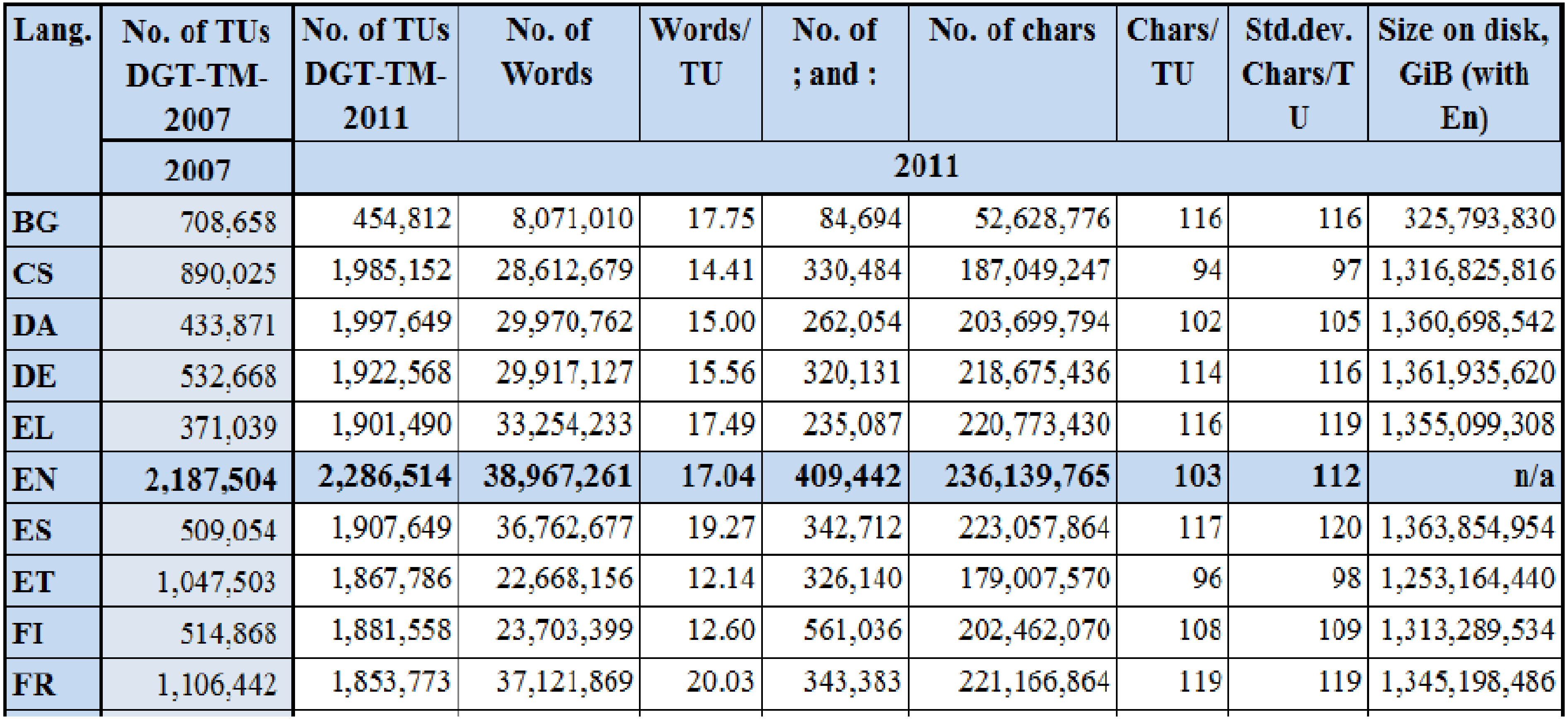
DGT-TM – some figures

Translation units	English	French	German
release 2007	2 187 504	1 106 442	532 668
release 2011	2 286 514	1 853 773	1 922 568
release 2012	322 377	273 961	284 072
release 2013	538 949	462 431	472 081
TOTAL	5 335 344	3 422 646	3 211 389

DGT-TM – some figures

Translation units	Greek	Swedish	Maltese
release 2007	371 039	555 362	1 021 855
release 2011	1 901 490	1 934 964	461 865
release 2012	285 483	283 589	263 804
release 2013	462 304	478 204	386 677
TOTAL	3 020 316	3 252 119	2 134 201

DGT-TM – Big data



Lang.	No. of TUs DGT-TM-2007	No. of TUs DGT-TM-2011	No. of Words	Words/ TU	No. of ; and :	No. of chars	Chars/ TU	Std.dev. Chars/TU	Size on disk, GiB (with En)
	2007	2011							
BG	708,658	454,812	8,071,010	17.75	84,694	52,628,776	116	116	325,793,830
CS	890,025	1,985,152	28,612,679	14.41	330,484	187,049,247	94	97	1,316,825,816
DA	433,871	1,997,649	29,970,762	15.00	262,054	203,699,794	102	105	1,360,698,542
DE	532,668	1,922,568	29,917,127	15.56	320,131	218,675,436	114	116	1,361,935,620
EL	371,039	1,901,490	33,254,233	17.49	235,087	220,773,430	116	119	1,355,099,308
EN	**2,187,504**	**2,286,514**	**38,967,261**	**17.04**	**409,442**	**236,139,765**	**103**	**112**	**n/a**
ES	509,054	1,907,649	36,762,677	19.27	342,712	223,057,864	117	120	1,363,854,954
ET	1,047,503	1,867,786	22,668,156	12.14	326,140	179,007,570	96	98	1,253,164,440
FI	514,868	1,881,558	23,703,399	12.60	561,036	202,462,070	108	109	1,313,289,534
FR	1,106,442	1,853,773	37,121,869	20.03	343,383	221,166,864	119	119	1,345,198,486

DGT public corpora – conditions of use

Conditions of use

- *Free of charge*
- *Exclusive property of EC.*
- *EC cedes its non-exclusive rights free of charge for all kinds of use which comply with the conditions laid down in the Commission Decision of 12 December 2011 on the re-use of Commission documents, (OJ L330).*
- *User has the obligation to state the source and the fact that the EC retains ownership of the data.*
- *Use of software in accordance with EUPL licence.*

DGT public corpora - access

- **"Language technology resources" pages of JRC:**
  http://ipsc.jrc.ec.europa.eu/?id=61
  together with other TM (EAC-TM, ECDC-TM)
  and resources (JRC-Names, JRC Eurovoc Indexer etc.)

- **"Open data portal" of the EU:**
  https://open-data.europa.eu/en/data
  only DGT-TM

More information

Reference publication *(focusing on 2011 release):*

"DGT-TM: A freely Available Translation Memory in 22 Languages"
by Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos and Patrick Schlüter

Contact persons:

For DGT-TM: Patrick.SCHLUTER@ec.europa.eu

For DGT-Acquis: Manuel.CARRASCO-BENITEZ@ec.europa.eu

For all language technology resources on the JRC site:
Ralf.STEINBERGER@jrc.ec.europa.eu

For DGT language applications:
Spyridon.PILOS@ec.europa.eu