SPA+ Page Count

Functional description

v3

Directorate for Support and Technological Services for Translation

Euramis Pre-Translation Unit

This document provides a description of the functional characteristics of the page count module used for the workload assessment calculation in the Directorate General for Translation. It provides definitions and descriptions of Gross Page Count, Net Page Count, Page Count Reports, as well as the rules applied in calculating the page counts.

Table of Contents

Background

General rules

The page count module performs the page count calculation and generates the page count reports. It is an integral part of the Safe Working Protocol Automation (SPA) pre-treatment system.

SPA+ page counting rules implement the decisions approved by the Working Group on Page Counting. They apply both for monolingual and multilingual documents, including mise-en-forme and relay translations of multilingual documents. Multilingual source documents require SPA+ to perform specific pre-processing steps in order to obtain the page count of the monolingual parts of such documents.

The page count module provides the calculation of gross pages (the total number of translation pages) and net pages (the number of translation pages taking into account the potentially reusable segments from the Euramis retrieval). The calculation is done based on the self-alignment and retrieval TMX files of the original document sent for translation. The self-alignment TMX file is used for the calculation of gross and net pages, whereas the retrieval TMX file is used to calculate the net pages only. The calculation method of the gross and net page count is described in the next sections of this document.

Special case

Following the changes adopted on 18 September 2012 to the page count of multilingual documents and their relay versions in the SPA+ packages for external contractors([1]), the page count of multilingual documents covers all source languages, including language combinations not covered by the contracts. If any part of the original multilingual document is to be translated via the relay language, the corresponding page count will be based on the page count of the original source language part in the multilingual document.

1 The change does not apply to the SPA+ packages for in-house translations where the page count of the relay translations is based on the relay original file.

Gross Page Count

Definition

Gross page count is the total number of source language pages to be translated.

Calculation

Gross pages are equal to the number of characters without spaces divided by 1500, excluding segments containing only numbers (e.g. cells in a table). The number of characters is calculated based on the content of the self-alignment TMX file of the input document received for translation.

Self-alignment TMX file

Self-alignment is the processing of a document using the Euramis Alignment in order to obtain a TMX file (**the self-alignment TMX file**) that contains the input document aligned against itself at sentence level (i.e. every source sentence is matched with itself). This process removes character strings which would not be included in an ordinary alignment, e.g. segments which only contain numerical characters and which need not be translated.

The content of the original document received for translation may differ from the content of its self-alignment due to the following facts:

- The self-alignment process takes account of the Euramis Alignment and document pre-processing rules whereby certain character strings and non-translatable text (such as codes, bullets, paragraph numbering, etc.) are deleted from the input document and thus excluded from the self-alignment TMX file.
- Segments that are not to be translated in the original document, such as attendance records, certain annexes or very long justifications, can be manually removed from the document prior to the self-alignment. This is done on a copy of the original file sent to translation. As a result, the removed segments are not included in the self-alignment TMX file and are not taken into account for the page count.
- In case of multilingual originals, as many input documents as source languages are sent for self-alignment. These input documents are the monolingual txt files (one per source language) split by the Language Detector from the multilingual original. **Invariable text**, such as the phrase 'Or.ll' – where 'll' stands for any of the official source languages of the EP, is excluded from the gross page count. Therefore, such phrases are not included in the input document sent for self-alignment.
- For XML4EP originals, the input document sent for self-alignment is a txt file containing the text extracted from the original file(s) based on extraction rules specific for the document type.

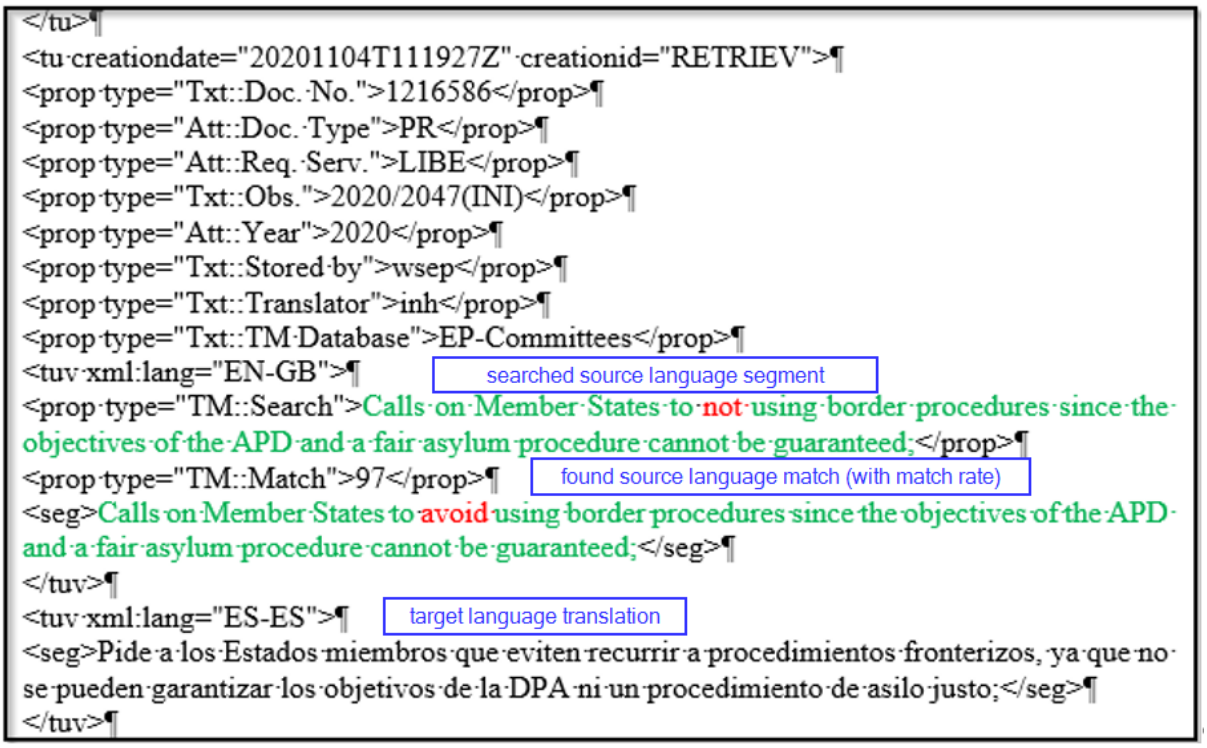
Net Page Count

Definition

Net page count represents the potential number of pages to be translated. It is estimated per individual target language into which a document is to be translated, taking account of the reusable segments included in the retrieval TMX file from Euramis.

Retrieval TMX file

Retrieval is a process whereby a source language document is segmented and the results are compared segment by segment with the contents of a maximum of 5 specified Euramis databases. The system then uses the metadata of the matches found in the source language to identify the corresponding segments in the specified target language(s) and create translation units which are included in a separate TMX file (**the retrieval TMX file**) for each specified source-target language pair. Only the three best matches above 65% from each database are kept.




```
</tu>¶
<tu creationdate="20201104T111927Z" creationid="RETRIEV">¶
<prop type="Txt::Doc. No.">1216586</prop>¶
<prop type="Att::Doc. Type">PR</prop>¶
<prop type="Att::Req. Serv.">LIBE</prop>¶
<prop type="Txt::Obs.">2020/2047(INI)</prop>¶
<prop type="Att::Year">2020</prop>¶
<prop type="Txt::Stored by">wsep</prop>¶
<prop type="Txt::Translator">inh</prop>¶
<prop type="Txt::TM Database">EP-Committees</prop>¶
<tuv xml:lang="EN-GB">¶
<prop type="TM::Search">Calls on Member States to not using border procedures since the objectives of the APD and a fair asylum procedure cannot be guaranteed;</prop>¶
<prop type="TM::Match">97</prop>¶
<seg>Calls on Member States to avoid using border procedures since the objectives of the APD and a fair asylum procedure cannot be guaranteed;</seg>¶
</tuv>¶
<tuv xml:lang="ES-ES">¶
<seg>Pide a los Estados miembros que eviten recurrir a procedimientos fronterizos, ya que no se pueden garantizar los objetivos de la DPA ni un procedimiento de asilo justo;</seg>¶
</tuv>¶
```

A translation unit in a retrieval TMX file showing the search string (source language segment), the found segment with the corresponding match rate, and the translation of the found segment.

The retrieval TMX is included in the pre-treatment package. It can be found in the Tmx project subfolder and it has been imported into the project *WTM*.

Calculation

The net page count relies on the potential re-use of segments included in the retrieval TMX file delivered by Euramis. It is calculated based on the content of the self-alignment TMX file and the retrieval TMX file.

The re-use figure, expressed in number of characters, is obtained by applying specific weighting/multiplication factors depending on the match rates indicated in the retrieval TMX file. The re-use figure is then subtracted from the gross number of characters calculated based on the self-alignment. The result is divided by 1500 in order to obtain the **number of pages to be paid**.

If no match is found for a segment from the self-alignment TMX file, leading bullets and paragraph numbering strings (e.g. A. or 1. or (i) or -) are removed and the match rate is re-attempted against the retrieval TMX file.

Weighting system

The matches from the retrieval TMX file are converted into standard pages using a formula based on the following **weighting system**:

- All document types, except AM

Match rate	Weighting	Comment
100% match	10%	The number of standard pages of the source text is multiplied by 0.10
95-99% match	15%	The number of standard pages of the source text is multiplied by 0.15
82-94% match	50%	The number of standard pages of the source text is multiplied by 0.50
0-81%match	100%	The number of standard pages of the source text is multiplied by 1

➢ AM documents

Match rate	Weighting	Comment
100% match	5%	The number of standard pages of the source text is multiplied by 0.05
95-99% match	10%	The number of standard pages of the source text is multiplied by 0.10
82-94% match	30%	The number of standard pages of the source text is multiplied by 0.30
0-81% match	100%	The number of standard pages of the source text is multiplied by 1

Repetitions

For **repetitions**, the first occurrence of any segment is counted once at the best match rate indicated in the retrieval TMX file. Subsequent occurrences are counted as 100% matches, depending on the document type.

Document type	Repetition weighting
All document types except AM	10%
AM documents	5%

Page count reports

Depending on whether the outsourced document is monolingual or multilingual, the provided pre-treatment package will include one or both types of page count reports described below.

Global page count report

This page count is included in the pre-treatment package (the main zip package) provided for multilingual documents.

It contains the following page count information:

- breakdown of gross and standard translation pages for each source language, including the count for source languages to be translated via relay;
- the total number of gross source pages;
- the total number of standard pages to be paid, equivalent to the net translation pages.

Directorate-General for Translation
Directorate A - Support and Technological Services for Translation

Global Page Count Report

Date: Thursday, 07 Nov 2024
Fdr: 6052746 / 1
Target language: EN

SOURCE LANGUAGE	PL	ES
Number of gross source pages	0.29	0.82
Standard pages to be paid	0.26	0.66

Total number of gross source pages: 1.11
Total number of standard pages to be paid: 0.92

This page count report contains the number of gross pages and the number of standard pages to be paid for all source languages of the original document, irrespective of whether the source language(s) is (/are) covered by the present contract or if you translate from a relay language. The report serves therefore both for the directly assigned source languages and for the languages assigned via relay.
(*) Source language translated from relay.

Source-target page count report

This page count report is included in the pre-treatment packages provided for monolingual documents, and also in the individual source-target language zip packages that are part of the main pre-treatment zip packages provided for multilingual originals.

The report contains the following page count information:

- a breakdown of the number of gross and net characters for each match rate range, with the corresponding multiplication factors;
- the total number of gross source pages;
- the total number of standard pages to be paid, equivalent to the net translation pages.

➢ <u>All document types, except AM</u>

Directorate-General for Translation
Directorate A - Support and Technological Services for Translation

Source-Target Page Count Report

Date: Wednesday, 06 Nov 2024
Fdr: 6052728 / 1
Source - Target Language: ES - LV

MATCH RATE	NUMBER OF GROSS CHARACTERS	PAYMENT	NUMBER OF STANDARD CHARACTERS
No match: 0%-81%	1134	100 %	1134
Match: 82%-94%	0	50 %	0
Match: 95%-99%	38	15 %	5
Match: 100%	68	10 %	6
TOTAL	**1240**		**1145**

Total number of gross source pages: 0.83
Total number of standard pages to be paid: 0.77
*** Number of pages = number of characters without spaces divided by 1500**

➢ AM documents

Европейски парламент Parlamento Europeo Evropský parlament Europa-Parlamentet Europäisches Parlament Euroopa Parlament Ευρωπαϊκό Κοινοβούλιο European Parliament Parlement européen Parlaimint na hEorpa Europski parlament Parlamento europeo Eiropas Parlaments Europos Parlamentas Európai Parlament Parlament Ewropew Europees Parlement Parlament Europejski Parlamento Europeu Parlamentul European Európsky parlament Evropski parlament Euroopan parlamentti Europaparlamentet

Directorate-General for Translation
Directorate A - Support and Technological Services for Translation

Source-Target Page Count Report

Date: Wednesday, 06 Nov 2024
Fdr: 6052681 / 2
Source - Target Language: EN - BG

MATCH RATE	NUMBER OF GROSS CHARACTERS	PAYMENT	NUMBER OF STANDARD CHARACTERS
No match: 0%-81%	1362	100 %	1300
Match: 82%-94%	40	30 %	12
Match: 95%-99%	274	10 %	27
Match: 100%	773	5 %	38
TOTAL	2449		1377

Total number of gross source pages: 1.64
Total number of standard pages to be paid: 0.92
* **Number of pages = number of characters without spaces divided by 1500**

Page count summary report

This detailed page count report can be provided upon request. It includes a segment-wise analysis of the document with the match rates found in the Euramis databases, as well as the corresponding weighting factors for the various match rates, and the resulting net values:

Segments

Id	Segment	Repetitions	Cleaning Attempted	Gross Chars	Match rate (%)	Payment (%)	Net Chars	Cumulative Net Chars
1	Shootingcanoeingsailing	0	true	25	0	100	25.00	25.00
2	ANNEX 2:	0	false	7	75	100	7.00	32.00
3	ANSWERS BY MAREK OPIOLA TO THE QUESTIONNAIRE	0	true	38	0	100	38.00	70.00
4	Do#wiadczenie zawodowe	0	false	21	100	20	4.20	74.20
5	Prosz# przedstawi# swoje do#wiadczenie zawodowe w dziedzinie finansów publicznych, zarówno w zakresie planowania bud#etu, realizacji bud#etu lub zarz#dzania bud#etem czy kontroli bud#etowej lub audytu.	0	false	178	100	20	35.60	109.80
6	Od pi#tnastu lat pracuj# w administracji pa#stwowej.	0	true	46	0	100	46.00	155.80
7	Wykonuj#c mandat Pos#a przez 5 kadencji Sejmu Rzeczypospolitej Polskiej, a nast#pnie sprawuj#c funkcj# Wiceprezesa Najwy#szej Izby Kontroli zdoby#em kompleksowe do#wiadczenie w procesie stanowienia prawa w dziedzinie finansów publicznych oraz w zakresie projektowania, kontroli i nadzoru nad wykonaniem bud#etu pa#stwa.	0	true	282	0	100	282.00	437.80
8	Podczas ponad 14-letniej dzia#alno#ci poselskiej zasiada#em w komisjach:	0	true	65	0	100	65.00	502.80
9	RegulaminowejObrony Narodowej oraz do Spraw S#u#b Specjalnych, gdzie pe#ni#em funkcje wiceprzewodnicz#cego oraz przewodnicz#cego.	0	true	118	0	100	118.00	620.80
10	Do moich priorytetowych zada# w komisjach nale#a#o m. in. coroczne przyjmowanie projektów bud#etów oraz sprawozda# z wykonania bud#etu pa#stwa, w tym tak istotnych instytucji, jak Kancelaria Prezydenta RP, Kancelarie Sejmu i Senatu, Krajowe Biuro Wyborcze, Ministerstwo Obrony Narodowej i s#u#by specjalne ds. bezpiecze#stwa pa#stwa, w tym bud#ety niejawne i wszystkie bud#ety o najwy#szej klauzuli niejawno#ci.	0	true	358	0	100	358.00	978.80
11	Te kadencje to 14 bud#etów pa#stwa, w tym 50 bud#etów niejawnych, 30 bud#etów w zakresie najwa#niejszych Kancelarii w kraju, 14 bud#etów w zakresie obronno#ci pa#stwa.	0	true	144	0	100	144.00	1122.80
12	Do#wiadczenie w projektowaniu oraz wykonaniu bud#etu pa#stwa na poziomie parlamentarnym przygotowa#o mnie do kolejnego etapu, jakim jest pe#niona obecnie przeze mnie funkcja	0	true	303	0	100	303.00	1425.80