Paper SD03

Automated RTF Table Translation
An Illustrated Approach

Qian Wang, MSD, Brussels, Belgium
Amit Bhaniani, MSD, London, United Kingdom
Carl Herremans, MSD, Brussels, Belgium

ABSTRACT
More and more countries request for HTA (Health Technology Assessment) dossiers to evaluate the added benefits of drugs compared to existing local standards of care in support of market access and price negotiations. While clinical trial results are typically reported in English, they need to be translated into local languages for HTA submissions. Manual translation from English to another language is not only time consuming but is also prone to errors and lacks traceability. Commonly available automated translation solutions are difficult to validate and lack traceability and reproducibility. This highlights the need for a more robust solution for table translation. This paper describes a SAS@ based tool to automate table translation from English to German. This tool can be extended to support translation to other languages. The translation process described here can also be implemented using programming languages other than SAS.

INTRODUCTION
Requests for HTA dossiers to evaluate added benefits of existing or new drugs are on an increase. A typical HTA dossier may include hundreds of tables containing clinical trial results. Most HTA agencies require the results to be presented in local language. As an example, German HTA dossiers mandate that all tables be prepared in German to support application for IQWiG opinion and G-BA decision. Short turnaround time and the need to produce all analyses in both English and local language add significant stress to the project teams working to deliver time-critical submission reports. In the absence of an automated solution, tables are translated manually by taking a copy of English version and replacing English text with translated text. This step is a bottleneck in the Analysis and Reporting (A&R) process and presents risk of incorrect copying of numbers, inaccurate and inconsistent translation of tables across multiple projects, significant use of company resources and lack of traceability and reproducibility. There is hence clearly a business need to automate table translation.

Most commonly available and used automated solutions amongst others are: Google Translate, Bing Translator, SYSTRAN and Moses (Wikipedia: Comparison of machine translation applications). These automated solutions, apart from challenges with respect to confidentiality of results and reproducibility, also lack control on exact terminology by the end user. Typically, as the HTA dossier develops, *consistency* and *precise choice of terminology* gets critical. Therefore, it is important that the translated results ensure this consistency and precise choice of terminology. This level of control is not addressed by the current readily-available software packages.

A robust automated translation program to prepare report-ready output tables with standardized layout will not only reduce time required to address critical HTA requests but also eliminate manual copying of numbers thereby minimizing translation errors, reducing use of resources and improving data traceability. Although standardizing layout for output tables takes time and effort, in the long run it helps to provide reproducible outputs and enables development of a limited number of common translation dictionaries which can be shared across multiple projects.

This paper provides a high-level introduction to our translation process and detailed description of each component using sample fake data. Maintenance and updates to the dictionary files are also explained. Finally, we discuss the key features of the automated translation process and conclude the paper with recommendations and areas for further improvements.

TRANSLATION COLLABORATION

Automated table translation is a joint effort requiring collaboration across different functional areas.

Figure 1 End to end table creation process flow

The end to end table translation process includes 3 main steps. The first step is the creation of English tables by the biostatisticians and the statistical programming team. Based on the English outputs, an English to German dictionary is built incrementally. Instead of manually translating each English table, in this automated process, the German dossier authors focus on how to maintain/enhance the dictionary to ensure consistency and reproducibility. The dictionary is then used by the programming team to recreate the same tables, this time in German. In this collaboration, each stakeholder makes use of their strength and expertise: the biostatistics and programming team on analysis and reporting, and the dossier authors on the local language, regulatory requirements and submission histories.

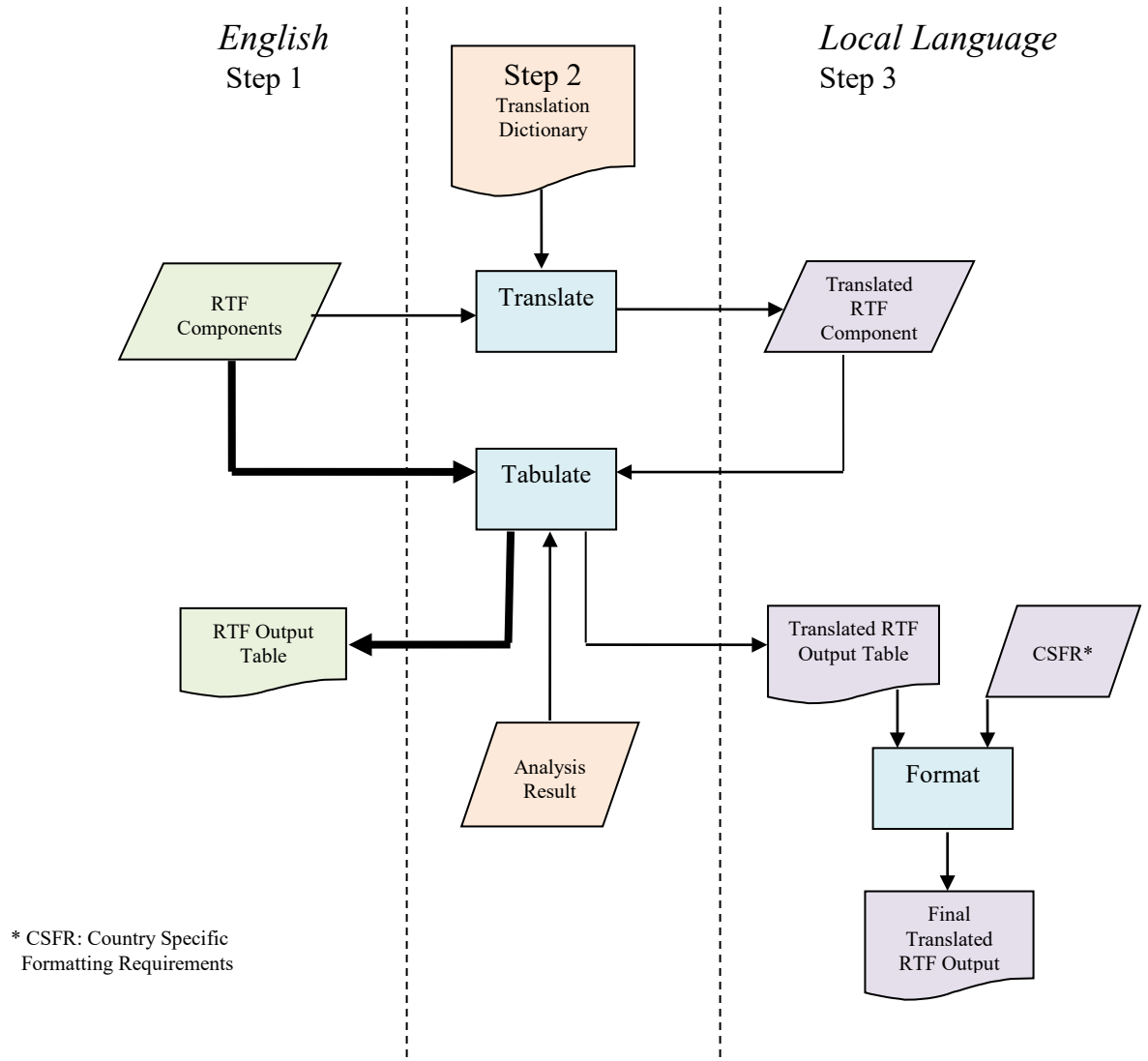

Figure 2 RTF table creation and translation process flow

The flow chart in Figure 2 illustrates the different components in Rich Text Format (RTF) table (Wikipedia - Rich Text Format) creation and translation. The core of the translation toolkit consists of two modules: Translate and Tabulate. The left side of the diagram with bold arrow lines describes how the English outputs are generated and the right side describes how these outputs are translated. Both English and translated RTF outputs are produced in a similar way through the same tabulation tool. The input to the "Tabulate" module includes the analysis results as well as the corresponding RTF components containing the information for the headers, titles, footers, and the text section of the table. When the tables are required in local language, the English RTF components are parsed through the "Translate" module to be translated using the dictionary files. The translated RTF components are then used instead of the English version for the "Tabulate" module to produce the translated RTF output. The toolkit also allows country/language specific RTF formatting to be added when needed.

The section below provides more detail on each component used in RTF table creation and translation process.

DESCRIPTION OF VARIOUS COMPONENTS OF THE TRANSLATION PROCESS

1. "Tabulate" Module
   %RTFTable macro (Eric Qi and Liping Zhang 2003) is used and described in this paper. However different tabulation solutions are possible in this framework. %RTFTable macro is a powerful SAS® tool to convert SAS® data sets into highly customized tables in RTF format. The macro provides flexibility to control the appearance of different components in the output table: title and footnotes, header, table body and its columns. As an example, users can easily specify the titles, and the footnotes to be used, the number of header rows, header content and formatting, the table column width and alignment. Special RTF symbols can also be incorporated anywhere in the output tables.

2. RTF output table
   Figure 3 shows a sample RTF table in English which will be used as an example to discuss the various steps involved in the table translation process. This table displays fake data.

Analyses of Overall Survival
(Intention-to-Treat Population)

Study: ABC	Treatment1			Treatment2			Treatment1 vs. Treatment2	
	N	Patients with Event n (%)	Median Time in Months [95 %-CI]	N	Patients with Event n (%)	Median Time in Months [95 %-CI]	Hazard Ratio [95 %-CI]	p-Value
Overall Survival	154	44 (28.6)	4.4 [0.7; 7.7]	151	64 (42.4)	1.14 [0.4; 2.15]	0.60 [0.41; 0.89]	0.010
Overall Survival - Adjusted	154	44 (28.6)	6.4 [4.2; 8.4]	151	46 (30.5)	0.75 [0.08; 1.13]	0.70 [0.46; 1.07]	0.101
Overall Survival - RPSFT	154	44 (28.6)	7.2 [6.5; 8.1]	151	45 (29.8)	0.65 [0.60; 0.71]	0.57 [0.32; 0.86]	0.010
Database Cutoff Date: 31DEC2099 Number of patients: intention-to-treat								

Figure 3 RTF output table example

3. RTF components
   Each standardized RTF table consists of four components as highlighted in Figure 4: (1) title and footnotes (2) header (3) first column and (4) body of the table. The contents and display of each component can be defined and provided as inputs for tabulation. These components are stored as SAS datasets to be translated.

(1)

Analyses of Overall Survival
(Intention-to-Treat Population)

(2) Study: ABC	Treatment1			Treatment2			Treatment1 vs. Treatment2	
	N	Patients with Event n (%)	Median Time in Months [95 %-CI]	N	Patients with Event n (%)	Median Time in Months [95 %-CI]	Hazard Ratio [95 %-CI]	p-Value
Overall Survival	154	44 (28.6)	4.4 [0.7; 7.7]	151	64 (42.4)	1.14 [0.4; 2.15]	0.60 [0.41; 0.89]	0.010
Overall Survival - Adjusted (3)	154	44 (28.6)	6.4 [4.2; 8.4]	151	46 (4)	0.75 [0.08; 1.13]	0.70 [0.46; 1.07]	0.101
Overall Survival - RPSFT	154	44 (28.6)	7.2 [6.5; 8.1]	151	45 (29.8)	0.65 [0.60; 0.71]	0.57 [0.32; 0.86]	0.010

Database Cutoff Date: 31DEC2099 (1)
Number of patients: intention-to-treat

Figure 4 RTF Components in a standard table

Figure 5 is a header component dataset corresponding to the sample English table. This table contains 2 headers. Header 1 consists of 4 columns, separated by vertical bars. Column 2, 3, and 4 are merged cell of several columns which is annotated with ^. Header 2 information is split into 3 observations each with a text length of 200 characters. RTF tags are included to define the text line wrap (\line) and font format (\b).

	NAME	OFFSET	VALUE
1	HEADER1	0	| \b | Treatment 1 ^^ | Treatment 2 ^^ | Treatment 1 vs. Treatment 2 ^ |
2	HEADER2	0	| \line Study: ABC \line \line |\line \line \line N|Patients \line with \line Event \line n (%)|Median \line Time in \line Months\line [95 %-CI] |\line \line \line N
3	HEADER2	200	|Patients \line with \line Event \line n (%)|Median \line Time in \line Months\line [95 %-CI] | \line \line Hazard Ratio\line [95 %-CI]|\line \line \line p-Value
4	HEADER2	400	\b0|
5	HEADERSIZE	0	4

Figure 5 RTF header component dataset (English)

4. Translation Dictionary

The translation dictionary files contain both English text and their translation in local language. Figure 6 is a partial header dictionary with English and its German translations for the example table in Figure 4. The dictionary file is built incrementally over time and is shared across studies. This ensures consistency and reproducibility of translation within and across projects. The RTF text formatting and text contents are separated as much as possible during the creation of the dictionary to minimize re-translating the same English text due to different RTF formats. In this example, table header “Patients with Event” is formatted with soft carriage return symbol \line. A different table with the same header may not need to break the text this way. The variation of RTF table header in this case will not be included in the dictionary as multiple entries. Only the English text “Patients with Event” is translated.

ENGLISH	TRANSLATION
Treatment	Behandlung
Patients with Event	Patienten mit Ereignis
Median Time in Months	Überlebenszeit in Monaten
N	N
n (%)	n (%)
[95 %-CI]	[95 %-KI]
Hazard Ratio	Hazard Ratio
p-Value	p-Wert
Treatment1	
Treatment2	
Treatment1 vs. Treatment2	

Figure 6 German RTF header translation dictionary

5. Translated RTF Components
   Using the dictionary files, RTF component datasets in English can be translated into local language of interest. The dataset in Figure 7 is the German version of the English header component dataset in Figure 5.

	VALUE	NAME	TRANSLATION
1		HEADER1	| \b | Behandlung1 ^ ^ | Behandlung2 ^ ^ | Behandlung1 vs. Behandlung2 ^ |
2	\line \line \line p-Value	HEADER2	| \line Studie: ABC \line \line | \line \line \line N | Patienten mit Ereignis \line n (%) | Mediane Überlebens-\line zeit in \line Monaten \line [95 %-KI] | \line \line \line N | Patienten mit Ereignis \line n (%) | Mediane Überlebens-\line zeit in \line Monaten \line [95 %-KI] | \line \line Hazard Ratio\line [95 %-KI] | \line \line \line p-Wert \b0 |
3		HEADER2A	|
4	Patients \line with \line Event \line n (%)	HEADER2B	| \line \line \line N | Patienten mit Ereignis \line n (%) |
5	Patients \line with \line Event \line n (%)	HEADER2C	| \line \line \line N | Patienten mit Ereignis \line n (%) |
6	\line \line \line p-Value	HEADER2D	| \line \line Hazard Ratio\line [95 %-KI] | \line \line \line p-Wert \b0 |

Figure 7 RTF header component dataset (German)

6. Country Specific Formatting Requirements
   RTF table translation not only translates text from one language to another, but also accommodates the country-specific formatting requirements. For example, in certain countries a comma (,) is used as a decimal separator instead of a dot (.), e.g. a value of 2.3 in the English table is expected to be printed as 2,3 in the German table. Figure 8 illustrates types of decimal separators used in different countries across the world.

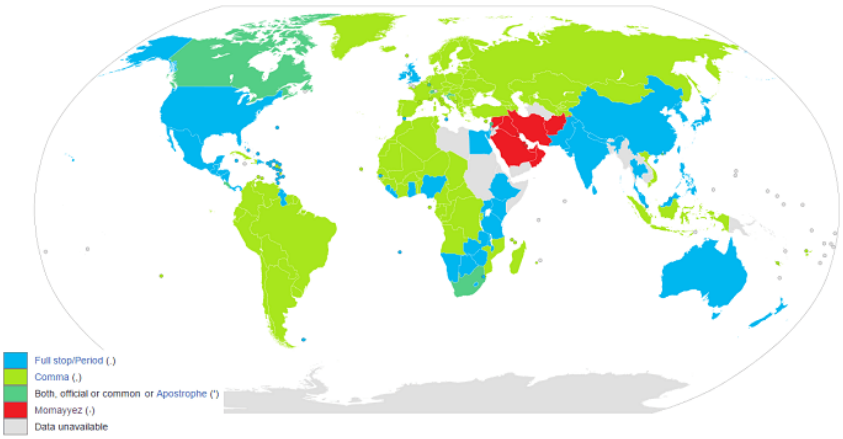

Figure 8 Decimal separators used around the world (Wikipedia: Decimal separator)

7. Translated RTF Output Table
The translated German output table corresponding to the sample English table is displayed in Figure 9. This table is mostly translated into German with some untranslated text (highlighted in red) and the country-specific decimal separator is also updated in this output table.

Analysen von Gesamtüberleben
(Intention-to-Treat Population)

Studie: ABC	Treatment1			Treatment2			Treatment1 vs. Treatment2	
	N	Patienten mit Ereignis n (%)	Mediane Überlebens-zeit in Monaten [95 %-KI]	N	Patienten mit Ereignis n (%)	Mediane Überlebens-zeit in Monaten [95 %-KI]	Hazard Ratio [95 %-KI]	p-Wert
Gesamtüberleben	154	44 (28,6)	4,4 [0,7; 7,7]	151	64 (42,4)	1,14 [0,4; 2,15]	0,60 [0,41; 0,89]	0,010
Gesamtüberleben - Angepasst	154	44 (28,6)	6,4 [4,2; 8,4]	151	46 (30,5)	0,75 [0,08; 1,13]	0,70 [0,46; 1,07]	0,101
Gesamtüberleben - RPSFT	154	44 (28,6)	7,2 [6,5; 8,1]	151	45 (29,8)	0,65 [0,60; 0,71]	0,57 [0,32; 0,86]	0,010
Datenschnitt: 31. Dec 2099 Anzahl der Patienten: Intention-to-treat								

Figure 9 Partially translated RTF output table in German

MAINTENANCE AND UPDATES TO THE TRANSLATION DICTIONARY

The development of the translation dictionary is an iterative process. The first iteration of the table translation process starts from an existing dictionary. This step, in addition to translating text from English to German, also generates a report which captures untranslated text i.e. English text without corresponding German text. This report is reviewed by the dossier authors who add new translations to the dictionary files and another iteration of the translation process is then performed. This process is repeated till all required English texts are translated into German.

For example, in the translated table in Figure 9, some text was left untranslated due to missing translations in the dictionary for “Treatment1”, “Treatment2” and “Treatment1 vs. Treatment2” in the table header. Figure 10 shows a partial output of the report where rows containing untranslated text are highlighted.

OUTPUT	SOURCE	ENGLISH_DISPLAY	TRANSLATED_DISPLAY	MATCH
SAMPLE1	HEADER	Treatment	Behandlung	Translated
SAMPLE1	HEADER	Treatment1		Not-Translated
SAMPLE1	HEADER	Treatment2		Not-Translated
SAMPLE1	HEADER	Treatment1 vs. Treatment2		Not-Translated
SAMPLE1	HEADER	Patients with Event	Patienten mit Ereignis	Translated
SAMPLE1	HEADER	Median Time in Weeks	Mediane Überlebenszeit in Monaten	Translated
SAMPLE1	HEADER	N	N	Translated
SAMPLE1	HEADER	n (%)	n (%)	Translated
SAMPLE1	HEADER	[95 %-CI]	[95 %-KI]	Translated
SAMPLE1	HEADER	Hazard Ratio	Hazard Ratio	Translated
SAMPLE1	HEADER	p-Value	p-Wert	Translated

Figure 10 Translation status report

The untranslated text can be easily identified using MATCH = ‘Not Translated’ condition. Translation for this text is updated to the translation dictionary as shown in Figure 11.

ENGLISH	TRANSLATION
Treatment	Behandlung
Patients with Event	Patienten mit Ereignis
Median Time in Months	Überlebenszeit in Monaten
N	N
n (%)	n (%)
[95 %-CI]	[95 %-KI]
Hazard Ratio	Hazard Ratio
p-Value	p-Wert
Treatment1	**Behandlung1**
Treatment2	**Behandlung2**
Treatment1 vs. Treatment2	**Behandlung1 vs. Behandlung2**

Figure 11 Updated German RTF header translation dictionary

Once the fallouts are added to the dictionary, the second iteration of the translation process successfully translates the sample English table (Figure 12).

Analysen von Gesamtüberleben
(Intention-to-Treat Population)

Studie: ABC	Behandlung1			Behandlung2			Behandlung1 vs. Behandlung2	
	N	Patienten mit Ereignis n (%)	Mediane Überlebens-zeit in Monaten [95 %-KI]	N	Patienten mit Ereignis n (%)	Mediane Überlebens-zeit in Monaten [95 %-KI]	Hazard Ratio [95 %-KI]	p-Wert
Gesamtüberleben	154	44 (28,6)	4,4 [0,7; 7,7]	151	64 (42,4)	1,14 [0,4; 2,15]	0,60 [0,41; 0,89]	0,010
Gesamtüberleben - Angepasst	154	44 (28,6)	6,4 [4,2; 8,4]	151	46 (30,5)	0,75 [0,08; 1,13]	0,70 [0,46; 1,07]	0,101
Gesamtüberleben - RPSFT	154	44 (28,6)	7,2 [6,5; 8,1]	151	45 (29,8)	0,65 [0,60; 0,71]	0,57 [0,32; 0,86]	0,010
Datenschnitt: 31. Dec 2099 Anzahl der Patienten: Intention-to-treat								

Figure 12 Fully translated RTF output table in German

The final translation step also generates the report (Figure 13) which confirms that no English text is left untranslated in the output tables.

OUTPUT	SOURCE	ENGLISH_DISPLAY	TRANSLATED_DISPLAY	MATCH
SAMPLE1	HEADER	Treatment	Behandlung	Translated
SAMPLE1	HEADER	Treatment1	**Behandlung1**	**Translated**
SAMPLE1	HEADER	Treatment2	**Behandlung2**	**Translated**
SAMPLE1	HEADER	Treatment1 vs. Treatment2	**Behandlung1 vs. Behandlung2**	**Translated**
SAMPLE1	HEADER	Patients with Event	Patienten mit Ereignis	Translated
SAMPLE1	HEADER	Median Time in Weeks	Mediane Überlebenszeit in Monaten	Translated
SAMPLE1	HEADER	N	N	Translated
SAMPLE1	HEADER	n (%)	n (%)	Translated
SAMPLE1	HEADER	[95 %-CI]	[95 %-KI]	Translated
SAMPLE1	HEADER	Hazard Ratio	Hazard Ratio	Translated
SAMPLE1	HEADER	p-Value	p-Wert	Translated

Figure 13 Updated translation status report

KEY FEATURES OF THE AUTOMATED TRANSLATION PROCESS

The motivation behind designing translation programs is to simplify and automate the manual process of output table translation, while addressing the key business need for report-ready output tables in local language with country-specific formatting. Some of the key features of our translation process are:

1. Ease of use: The translation process relies on the use of dictionary files which can be maintained as simple text files, SAS@ datasets or spreadsheets, making them easier to create, maintain and update.

2. Language independent: The translation process is language independent. Using an appropriate dictionary, this process can be used to translate outputs to any language of interest. Also, the translation process can be implemented using programming languages other than SAS as well.

3. Consistent, reproducible and traceable results: Since the translation dictionary can be shared across multiple projects, all translated tables would have identical translations and formatting specifications. This provides consistent and reproducible output tables. Adding the versioning system to the dictionary files would provide traceability by matching translated output tables to the dictionary files used for their translation

4. Modularity: The functionality to add country-specific formatting is modular in nature and can be customised for the country and language required. New formatting attributes can be implemented as add-ons to the existing tool to provide country-specific, translated, report-ready and immutable output tables.

5. Integration with the existing A&R processes: The translation process is flexible enough to mould around existing A&R processes as the same Tabulate module is used in the creation of both English and translated tables.

RECOMMENDATIONS AND LIMITATIONS

The success of the translation process just like any other programming initiative comes from its regular usage and feedback from end-users. Based on our experience of designing and implementing the translation process we can make a few recommendations: (1) Dictionary files should be saved and shared across projects to promote standard and consistent translation. (2) Including version-control information to the dictionary files ensures traceability of the translated outputs.

While the translation tool has greatly increased the quality and efficiency, there is room for improvement. Some of the limitations observed are: (1) The current translation tool does not translate the text within the table body (apart from the first column). A table with more sophisticated layout, e.g. multiple sections with different section headers in the body of the table, will not be completely translated. (2) The translation dictionary is manually maintained which is prone to errors and requires its own validation steps. More advanced data entry tool can be developed.

ACKNOWLEDGEMENTS

The authors would like to thank MSD BARDS HTA and MSD Sharp & Dohme GmbH HTA Germany colleagues for their contribution and collaboration towards automating and standardizing the translation process.

REFERENCES

1. Wikipedia contributors. (2019, February 25). Comparison of machine translation applications. In Wikipedia, The Free Encyclopedia. Retrieved March 7, 2019, from https://en.wikipedia.org/w/index.php?title=Comparison_of_machine_translation_applications&oldid=885009170
2. Eric Qi, Merck & Co; Liping Zhang (2003) %RTFTable - A Powerful SAS® Tool to Produce Rich Text Format Tables. PharmaSUG 2003. Retrieved March 7, 2019, from http://www.lexjansen.com/pharmasug/2003/TechnicalTechniques/tt038.pdf
3. Wikipedia contributors. (2019, February 25). Rich Text Format. In Wikipedia, The Free Encyclopedia. Retrieved 13:59, March 7, 2019, from https://en.wikipedia.org/w/index.php?title=Rich_Text_Format&oldid=885066330
4. Wikipedia contributors. (2019, March 6). Decimal separator. In Wikipedia, The Free Encyclopedia. Retrieved March 7, 2019, from https://en.wikipedia.org/w/index.php?title=Decimal_separator&oldid=886405958

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Amit Bhaniani
MSD
Pancras Rd 2, Kings Cross,
London N1C 4AG [United Kingdom]
Tel.: +441992452504
E-mail: amit.bhaniani@merck.com
Web: http://ww.msd.com

Qian Wang
MSD
Clos du Lynx 5, B-1200 Brussels [Belgium]
Tel: +3227766303
E-mail: qian_wang@merck.com
Web: http://www.msd.com

Carl Herremans
MSD
Clos du Lynx 5, B-1200 Brussels [Belgium]
Tel: +3227766305
E-mail: carl.herremans@merck.com
Web: http://www.msd.com