# USE OF DICTIONARIES IN AUTOMATED RTF TABLE TRANSLATION

Qian Wang[1], Ligia Adamska[2]

Biostatistics and Research Decision Sciences, MSD – (Brussels[1], London[2])


PhUSE SDE @ MSD – Brussels

Sep 21st 2018

# Introduction

- Both the local regulatory submission and the HTA (Health Technology Assessment) submission typically require the clinical trial results to be reported in local language instead of English.

- The HTA dossier authors are responsible for the translation of English text in the RTF tables to the local language.

- Previously, all RTF tables were translated manually to local language by MSD subsidiary.

- With 100s or 1000s of tables needing translation each year,

  an automated tool was much desired.

MSD
INVENTING FOR LIFE

# Thinking behind dictionary driven automated translation process…

- As the result tables only contain a limited vocabulary, automated RTF translation tool can be developed to improve translation efficiency.

Titles and footnotes

*Table 15:*
*Time to event for subgroup analyses of single endpoint, single study*

| study: <name> | < intervention > | | | < comparator > | | | < intervention > vs. < comparator > | | |
|---|---|---|---|---|---|---|---|---|---|
| < endpoint > Header | N[a] | patients with event n (%) | median time in < unit time > [95%-CI] | N[a] | patients with event n (%) | median time in < unit time > [95%-CI] | hazard ratio[b] [95 %-CI] | p-value[b,c] | p-value for interaction test ($I^2$) |
| < subgroup 1 > | | | | | | | | | |
| < characteristic 1a > First column | X | X (X,X) | X,X [X,X; X,X ] | X | X (X,X) | X,X [X,X; X,X ] | X,XX [X,XX; X,XX] | X,XXX | X,XXX (X,XX%) |
| < characteristic 1b > | X | X (X,X) | X,X [X,X; X,X ] | X | X (X,X) | X,X [X,X; X,X ] | X,XX [X,XX; X,XX] | X,XXX | |
| < subgroup 2 > | Numerical values | | | | | | | | |
| < characteristic 1a > | X | X (X,X) | X,X [X,X; X,X ] | X | X (X,X) | X,X [X,X; X,X ] | X,XX [X,XX; X,XX] | X,XXX | X,XXX (X,XX%) |
| < characteristic 1b > | X | X (X,X) | X,X [X,X; X,X ] | X | X (X,X) | X,X [X,X; X,X ] | X,XX [X,XX; X,XX] | X,XXX | |

a: number of patients: < explanation of analysis population (e.g. full analysis set, all-patients-as-treated, intention-to-treat, etc. >
b: < explanation of model (z.B. Cox proportional hazard model, stratified by ..., etc)
c: < explanation (e.g. by Wald, CSZ-method, Fisher's exact test, etc. >
CI: confidence interval, < definition of abbreviations >

Titles and footnotes

MSD
INVENTING FOR LIFE

# Results

## English version

Subject Medical History Conditions
(Incidence >0% in One or More Treatment Groups)
Population with TPS≥50%
(Intention-to-Treat Population)

| Study: [redacted] | [redacted] | | Standard Treatment | |
|---|---|---|---|---|
| | n | (%) | n | (%) |
| Subjects in population[b] | 64 | | 65 | |
| With one or more conditions[c,d] | 64 | (100.0) | 64 | (98.5) |
| With no conditions | 0 | (0.0) | 1 | (1.5) |
| **Blood and lymphatic system disorders** | **21** | **(32.8)** | **13** | **(20.0)** |
| Anaemia | 16 | (25.0) | 9 | (13.8) |
| Anaemia of malignant disease | 1 | (1.6) | 0 | (0.0) |
| Bone marrow oedema | 0 | (0.0) | 1 | (1.5) |
| Febrile neutropenia | 0 | (0.0) | 1 | (1.5) |
| Leukopenia | 1 | (1.6) | 1 | (1.5) |
| Lymph node pain | 0 | (0.0) | 1 | (1.5) |
| Lymphadenopathy | 3 | (4.7) | 0 | (0.0) |
| Lymphadenopathy mediastinal | 1 | (1.6) | 0 | (0.0) |
| Macrocytosis | 1 | (1.6) | 0 | (0.0) |
| Neutropenia | 1 | (1.6) | 2 | (3.1) |
| Normochromic normocytic anaemia | 2 | (3.1) | 0 | (0.0) |
| Thrombocytopenia | 1 | (1.6) | 1 | (1.5) |
| **Cardiac disorders** | **11** | **(17.2)** | **15** | **(23.1)** |
| Aortic valve disease | 0 | (0.0) | 1 | (1.5) |
| Aortic valve incompetence | 0 | (0.0) | 1 | (1.5) |
| Aortic valve sclerosis | 0 | (0.0) | 1 | (1.5) |
| Arrhythmia | 2 | (3.1) | 0 | (0.0) |
| Arteriosclerosis coronary artery | 0 | (0.0) | 1 | (1.5) |
| Atrial fibrillation | 2 | (3.1) | 2 | (3.1) |
| Atrial flutter | 1 | (1.6) | 0 | (0.0) |
| Atrioventricular block first degree | 0 | (0.0) | 1 | (1.5) |
| Bradycardia | 1 | (1.6) | 3 | (4.6) |
| Brugada syndrome | 0 | (0.0) | 1 | (1.5) |
| Bundle branch block left | 1 | (1.6) | 0 | (0.0) |
| Bundle branch block right | 0 | (0.0) | 2 | (3.1) |
| Cardiac disorder | 1 | (1.6) | 1 | (1.5) |
| Cardiac failure | 1 | (1.6) | 0 | (0.0) |
| Cardiac flutter | 1 | (1.6) | 0 | (0.0) |
| Congestive cardiomyopathy | 1 | (1.6) | 0 | (0.0) |
| Coronary artery disease | 0 | (0.0) | 1 | (1.5) |

## German translation

Subject Medical History Conditions
(Incidence >0% in One or More Treatment Groups)
Population with TPS≥50%
(Intention-to-Treat Population)

| Studie: [redacted] | [redacted] | | Standardtherapie | |
|---|---|---|---|---|
| | n | (%) | n | (%) |
| Anzahl an Patienten[b] | 64 | | 65 | |
| Mit einer oder mehreren Erkrankungen[c,d] | 64 | (100,0) | 64 | (98,5) |
| Ohne Erkrankung | 0 | (0,0) | 1 | (1,5) |
| **Erkrankungen des Blutes und des Lymphsystems** | **21** | **(32,8)** | **13** | **(20,0)** |
| Anaemie | 16 | (25,0) | 9 | (13,8) |
| Anaemie aufgrund einer malignen Erkrankung | 1 | (1,6) | 0 | (0,0) |
| Knochenmarkoedem | 0 | (0,0) | 1 | (1,5) |
| Febrile Neutropenie | 0 | (0,0) | 1 | (1,5) |
| Leukopenie | 1 | (1,6) | 1 | (1,5) |
| Lymphknotenschmerzen | 0 | (0,0) | 1 | (1,5) |
| Lymphadenopathie | 3 | (4,7) | 0 | (0,0) |
| Mediastinale Lymphadenopathie | 1 | (1,6) | 0 | (0,0) |
| Makrozytose | 1 | (1,6) | 0 | (0,0) |
| Neutropenie | 1 | (1,6) | 2 | (3,1) |
| Anaemie normochrom normozytaer | 2 | (3,1) | 0 | (0,0) |
| Thrombozytopenie | 1 | (1,6) | 1 | (1,5) |
| **Herzerkrankungen** | **11** | **(17,2)** | **15** | **(23,1)** |
| Aortenklappenvitium | 0 | (0,0) | 1 | (1,5) |
| Aortenklappeninsuffizienz | 0 | (0,0) | 1 | (1,5) |
| Aortenklappensklerose | 0 | (0,0) | 1 | (1,5) |
| Arrhythmie | 2 | (3,1) | 0 | (0,0) |
| Arteriosklerose der Koronararterie | 0 | (0,0) | 1 | (1,5) |
| Vorhofflimmern | 2 | (3,1) | 2 | (3,1) |
| Vorhofflattern | 1 | (1,6) | 0 | (0,0) |
| Atrioventrikulaerer Block ersten Grades | 0 | (0,0) | 1 | (1,5) |
| Bradykardie | 1 | (1,6) | 3 | (4,6) |
| Brugada-Syndrom | 0 | (0,0) | 1 | (1,5) |
| Linksschenkelblock | 1 | (1,6) | 0 | (0,0) |
| Rechtsschenkelblock | 0 | (0,0) | 2 | (3,1) |
| Herzerkrankung | 1 | (1,6) | 1 | (1,5) |
| Herzinsuffizienz | 1 | (1,6) | 0 | (0,0) |
| Herzflattern | 1 | (1,6) | 0 | (0,0) |
| Kongestive Kardiomyopathie | 1 | (1,6) | 0 | (0,0) |
| Koronare Herzerkrankung | 0 | (0,0) | 1 | (1,5) |
| Stenose der Koronararterie | 1 | (1,6) | 0 | (0,0) |

MSD
INVENTING FOR LIFE

# THANK YOU!

Please come and see our Poster!

# Use of Dictionaries in Automated RTF Table Translation

Qian Wang[1], Ligia Adamska[2]
Biostatistics and Research Decision Sciences, MSD Europe – (Brussels[1], London[2])

## Introduction

Both the local regulatory submission and the HTA (Health Technology Assessment) submission typically require the clinical trial results to be reported in local language instead of English. For a pharmaceutical enterprise operating on a global scale, manual translation of tabulated results in Rich Text Format (RTF) from English to another language is labour intensive, time consuming and prone to errors. As those result tables only contain a limited vocabulary, automated RTF translation tool can be developed to improve translation efficiency. Such tools make use of dictionaries, which provide the mapping between English and the local language. The quality of translation is largely driven by the quality of the dictionary.

This presentation describes a SAS® based tool and its dictionaries for automated RTF table translation using German language as an example. The dictionaries are constructed in a way to ensure consistency and traceability in the translated outputs, while allowing flexibility for specific translation.

## Methods

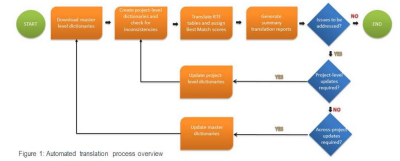


Figure 1: Automated translation process overview

### Process Overview

The automated translation process is detailed in *Wang&Bhaniani, 2018*. The translation is dictionary driven. The dictionary files provide the mapping between the English and German language and they are constructed in a way to ensure consistency and traceability in the translated outputs, while allowing flexibility for specific translation. The dictionaries consist of 4 documents, corresponding to different sections in an RTF table: MedDRA Preferred terms displayed in the first column of the table, terms related to table titles and footnotes, terms related to table headers, and terms related to the first column of an RTF table (see Figure 2 for a table schema)

### Master Dictionaries

The master dictionaries are documents containing standard terms used across projects that have been centralised and are manually maintained by country level HTA dossier authors to ensure appropriate translation. Master dictionaries contain *wildcards* to reduce the amount of unnecessary translation work due to project specific formatting (e.g. N=@@@ instead of N=123).

### Project –Level Dictionaries and PRX Language

Project-level dictionaries are project specific. They are automatically generated by replacing the *wildcards* in the master dictionary files with the project specific formatting. Perl Regular Expression (PRX) language is employed for the pattern detection and text replacement. (e.g. replacing N=@@@ with N=123). Project dictionaries can also be updated manually based on a project-specific requirement (e.g. \line). Inconsistency check is performed between project-level and master dictionary files to ensure consistency in the translated text.

### Translation Output and Best Match

During translation process a summary report is generated (see Figure 3 for an example) and translation status (TRANSLATED, NOT TRANSLATED and BEST MATCH) is assigned to each English text segment to aid the user in understanding the current status and identifying the action needed. A BEST MATCH translation is assigned when the translation is partially known albeit the English term in output table does not match exactly the term in the dictionary (e.g. a variation of the same footer). By reviewing the status report, the users can then decide whether the master or project-level dictionaries should be updated as seen in the automated translation process schema in Figure 1. The level of update depends on its scope. If a term is missing from the dictionary, then a request should be made to the HTA dossier authors for this term to be added to the master dictionary. If an already existing term needs an update to translation or the English text and this update is to be used globally (i.e. across projects), then the master dictionaries should be updated. Otherwise, if the update is project specific, only the project-level dictionaries should be updated by the user.

## Results

The automated translation process was tested on tabular outputs for the German dossier using SAS® 9.3 on UNIX platform. A total of 263 RTF tables were successfully translated from English to German language. The process is iterative and it took 2 full runs to complete. The time taken for each run was 1h40mins. For the tested RTF tables, only footnotes, headers and 1st column needed translating. Titles did not need translation. (see Figure 2 below).

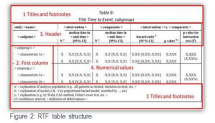


Figure 2: RTF table structure

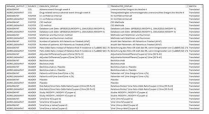

Figure 3: Summary translation report

### Dry Run

Following the 1st run, a summary translation report indicated that 1 term (other than title) was not translated and 5 terms were assigned a BEST MATCH value. All other terms were translated correctly. For the one term that was not translated, the decision was that the master dictionary should be updated. For the 5 terms with BEST MATCH, for 3 of those English tables needed to be updated to match the term in dictionary and for the other 2 project-level dictionaries were updated.

### Live Run

The 2nd run of translation got all the necessary fields in the table successfully translated to German language (refer to Figure 3 for a snapshot of summary report and to Figure 4 for an example of a translated table).

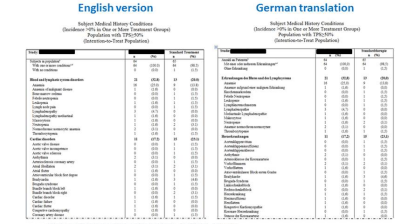


Figure 4: An example of RTF table translated to German language

## Conclusions

In this presentation we have showed the automated, dictionary driven RTF table translation process and how it was used to successfully translate RTF tables to German language. The dictionaries, that are pivotal to the translation process, are constructed in a way to ensure consistency and traceability in the translated outputs, while allowing flexibility for specific translation by employing the PRX language. By limiting the human element in a translation process, it makes the tool more robust, accurate and less costly as compared to manual translation.

Future work will focus on limiting the number of manual steps that are still required in the translation process, making the process more efficient by reducing the processing time and possibly expand to other local languages.

**References:**

Qian Wang & Amit Bhaniani , 2018, RTF Table Translation of clinical trial results: An automated approach, PhUSE SDE 2018

**Contact information**

qian.wang@merck.com

ligia.adamska@merck.com

MSD