*MEDINFO 2017: Precision Healthcare through Informatics*
*A.V. Gundlapalli et al. (Eds.)*

*doi:10.3233/978-1-61499-830-3-848*

# Comparison of Three English-to-Dutch Machine Translations of SNOMED CT Procedures

**Ronald Cornet[a,b], Carly Hill[a], Nicolette de Keizer[a]**

*[a] Department of Medical Informatics, Academic Medical Center – University of Amsterdam, Amsterdam Public Health Research Institute, The Netherlands*
*[b] Department of Biomedical Engineering, Linköping University, Linköping, Sweden*

**Abstract**

*Dutch interface terminologies are needed to use SNOMED CT in the Netherlands. Machine translation may support in their creation. The aim of our study is to compare different machine translations of procedures in SNOMED CT. Procedures were translated using Google Translate, Matecat, and Thot. Google Translate and Matecat are tools with large but general translation memories. The translation memory of Thot was trained and tuned with various configurations of a Dutch translation of parts of SNOMED CT, a medical dictionary and parts of the UMLS Metathesaurus. The configuration with the highest BLEU score, representing closeness to human translation, was selected. Similarity was determined between Thot translations and those by Google and Matecat. The validity of translations was assessed through random samples. Google and Matecat translated similarly in 85.4% of the cases and generally better than Thot. Whereas the quality of translations was considered acceptable, machine translations alone are yet insufficient.*

***Keywords:***

SNOMED CT; Natural Language Processing

## Introduction

The use of SNOMED CT in electronic health systems is growing [1]. Recording clinical data using SNOMED CT helps to uniformly describe medical data, which enables data reuse such as data analysis, auditing quality of care and decision support.

SNOMED CT is officially released in English and Spanish. Before SNOMED CT can be used in clinical practice in other languages, a translation needs to be made, or interface terminologies need to be created. SNOMED CT has been fully translated in Denmark and in Sweden. A Canadian French translation is ongoing. Similar to other countries [2], in the Netherlands a partial translation of SNOMED CT is undertaken. Such translations generally provide one term for each concept, adhering to the strict translation guidelines of SNOMED International[1]. However, they do not necessarily contain the synonyms used in clinical practice, and these synonyms may not adhere to the translation guidelines. Interface terminologies provide a close-to-user description for concepts, generally covering a part of SNOMED CT, e.g., diagnoses or procedures. Also for Dutch healthcare, one or more Dutch interface terminologies for SNOMED CT need to be made available. Descriptions for a part of the diagnoses form the Dutch interface terminology "Diagnoses thesaurus", which is maintained by Dutch Hospital Data (DHD) [2]. The next step is to start creation of an interface terminology for procedures. Recording of procedures is an essential part of clinical documentation and serves, once standardized, many data reuse purposes. These include calculation of quality indicators and reimbursement.

SNOMED CT contains more than 55,000 procedures, each described by one or more English descriptions. Manual translation of these descriptions requires a lot of time and resources. If computers are used to make initial translations, terms only have to be validated, which may save a lot of time. Machine translation has already been used for translating SNOMED CT in Spanish, Swedish and French. To make the Spanish version, prefixes, suffixes and roots of terms were used to make an automated proposal [3]. In Sweden, mappings to other already translated terminologies were used [4]. In France, lexical methods and mapping to the Unified Medical Language System Metathesaurus (UMLS Metathesaurus) were used [5].

Above methods were well evaluated, but not much research has been performed on already available translation tools. Hence, in this study, we assess the quality of the translation of descriptions of concepts in the procedures hierarchy of SNOMED CT from English to Dutch. We compare generic translation tools and a tool with a translation memory that was specifically trained and tuned for this purpose.

The first generic tool is Google Translate[3], the second generic tool is Matecat[4] [6], and the third tool is Thot[5] [7], a toolkit for statistical machine translation, which requires training and tuning of the translation memory.

The hypothesis is that translations that are the same among the three methods are of better quality, because three different translation systems have found the same result. We furthermore hypothesize to get less but better translations with Thot as this tool is trained with terms from a medical background.

## Materials & Methods

### Tools

Machine translation was performed using Google Translate, Matecat, and Thot.

Google Translate is a widely used translation tool. The translator is a black box with a large translation memory. It

[1] http://www.ihtsdo.org/resource/resource/9

[2] https://www.dhd.nl/klanten/producten-diensten/diagnosethesaurus/Paginas/Diagnosethesaurus.aspx
[3] https://translate.google.com/
[4] https://www.matecat.com/
[5] http://daormar.github.io/thot/

has previously been used for translation of SNOMED CT, e.g., in the research of Schulz et al. [2].

Matecat [6], is a tool that has its own translation memory. One difference with Google Translate is that one can select the subject of the text to translate. This context might help with choosing the right translation for a term. We selected 'Medical/Pharmaceutical' as the subject of the Matecat translation.

Thot is a phrase-based tool [7], i.e., it is usually trained by providing pairs of translations of phrases. In this research, these pairs consisted of Dutch and English phrases, predominantly noun phrases. Apart from providing the basic functionality for preparing, training and translating files, this tool also provides interactive translation possibilities. This means the system provides functionality for the user to have an influence on the final translations. For training, incremental learning is used to result in better language models.

**Sources for Training, Tuning and Testing Thot**

We used three resources to train, tune and test Thot. First, we used Dutch-English translation pairs from the UMLS Metathesaurus. Second, an existing but partial Dutch translation of SNOMED CT that has been developed in the joint Dutch-Belgian efforts to develop Dutch interface terminologies. Third, a Dutch-English medical dictionary, Springer Groot Medisch Woordenboek [8].

Thot tests were compared by means of the BLEU (bilingual evaluation understudy) score generated by Thot [9]. This score measures the closeness to a human translation in a range from zero to one, with one being a perfect human translation. The configuration that gave the best test result after training and tuning was used for the final translation of the SNOMED CT descriptions. This resulted in the Thot translation, which was compared to the Google Translate and Matecat translations.

**Filtering non-translations**

The fully specified names of all procedure concepts of the January 2016 release of SNOMED CT were selected. Each tool created a translation for each of the fully specified names, after removal of the semantic tags "(procedure)" and "(regime/therapy)".

Tools may fail to translate some words, resulting in untranslated English words in a "Dutch" translation. For example, if "Urinary undiversion" is translated as "urine undiversion," the second word is not a Dutch word. However, if "Open drainage of liver" is translated as "Open drainage van de lever," this is perfectly correct.

Hence, before comparing the three translations the amount of remaining English words was assessed. This was done by a Java program that checked if the words were in a Dutch list of words. This list consisted of generic Dutch words[6] merged with the words from the Dutch file we used to train Thot for domain-specific words. Translations containing more non-Dutch words than Dutch words were deleted. The amounts and percentages of included terms were calculated for each of the three tools, and the results were compared with McNemar tests.

**Validity of translations**

For each SNOMED CT concept we constructed a set of translations, as shown in Table 1. Sets with three equal translations, two equal translations and all different translations were created. We calculated the percentage of sets with at least one exactly similar translation by another method.

From each of the sets, a sample of 100 English terms was selected, with one, two and three different Dutch translations respectively. These samples of the translations were checked on validity: two reviewers (RC & CH) assessed whether the translations were well-formed Dutch noun-phrases reflecting the meaning of the English description. The meaning of the translation could be different, for example due to translation of an English homonym in the wrong context, such as vessels as ships ("schepen") (see Table 1), or stool as furniture ("kruk"). The reviewers graded using marks from zero to three, with zero and one being not acceptable and two and three being acceptable as a good translation. If an English word is recognizable as a Dutch word, the term will get 2 points, otherwise not more than 1 point. Determiners are not considered; wrong spacing costs a point, as Dutch is a language in which words are combined, e.g., "Mouth reconstruction" is "mondreconstructie", not "mond reconstructie".

A translation was considered acceptable when it was recognizable as a translation that covers the meaning of the English term. Average marks and 95% confidence intervals (CI) were calculated for the different samples and reviewers. Also, total weighted averages were calculated for the three tools. The percentages of the translations that were considered acceptable were calculated. For the samples containing two or three different translations, the reviewers assessed which translation they considered the best.

*Table 1 – Examples of three sets of translations made by the tools.*

| | English term | Google | Matecat | Thot |
|---|---|---|---|---|
| **1** | Oral sedation | orale sedatie | orale sedatie | orale sedatie |
| **2** | CT of pancreas | CT van de alvleesklier | CT van de alvleesklier | ct pancreas |
| **3** | Ultrasound scan of abdominal vessels | Echografie van abdominale schepen | Echografie van de buik vaartuigen | echografie abdominale bloedvaten |

## Results

**Training Toth**

Table 2 shows the results of the different Thot translations with their BLEU scores. The configuration with the highest score was used to compare to the general translation tools. This turned out to be a combination of Dutch terms (coming from other thesauri like ICD, ICPC) that UMLS relates to SNOMED CT concepts, terms from a Dutch-English medical dictionary, and terms from a Dutch translation of parts of SNOMED CT.

**Filtering non-translations**

SNOMED CT contained 54419 procedure concepts.

Sets of translations (i.e., all translations for a concept) were excluded if any of their translation contained more English words than Dutch words. Table 3 shows the number and percentage of included translations for each of the tools used.

[6] http://www.opentaal.org/bestanden.html

*Table 2 – BLEU scores generated by Thot for different training and tuning configurations.UMLS CT = SNOMED CT concepts from UMLS; UMLS procedures = SNOMED CT procedures from UMLS; Dict = Medical Dictionary; Trans = partial translation of SNOMED CT in Dutch.*

| Terms in training and tuning files | BLEU score |
|---|---|
| UMLS CT + Dict + Trans (tuned with procedures) | 0.596 |
| UMLS CT + Dict (tuned with procedures) | 0.430 |
| UMLS CT (tuned with procedures) | 0.427 |
| UMLS procedures | 0.407 |
| UMLS CT | 0.357 |

*Table 3 – Translation with amount and percentage of included terms. Total amount of terms to translate was 54419.*

| Translation | Number and percentage of translations included |
|---|---|
| Google | 52399 (96.3%) |
| Matecat | 52414 (96.3%) |
| Thot | 52685 (96.8%) |

*Table 4 – Amount of translations after checking and comparing files. Total number of included terms was 50838.*

| Equal translations | Number (percentage) of terms | |
|---|---|---|
| All translations equal | 1548 | (3.0%) |
| Two different translations | 42132 | (82.9%) |
| - Google & Matecat vs. Thot | 41865 | |
| - Google & Thot vs. Matecat | 180 | |
| - Matecat & Thot vs. Google | 87 | |
| All translations different | 7158 | (14.1%) |

There was no practical difference in the number of translated terms, whereas Thot translated ($p < 0.001$) significantly more terms into Dutch than Google and Matecat, which were not significantly different ($p = 0.535$). In all translations, less than 4% of the translations were rejected. Excluding all preferred terms for which one or more of the tools didn't provide a translation resulted in a set of 50838 terms. This set was used for further analysis.

For these terms we compared whether one or more of the tools provide the same translation. The results of this analysis are shown in Table 4.

Table 4 shows that full agreement between the three tools is much less common than full disagreement, and that Google and Matecat agree most of the time, in 43413 (85.4%) of the cases.

**Validity of translations**

We selected 100 sets from 1548 with all translations equal, i.e., 100 translated terms; 100 sets from 41865 where Google and Matecat agreed, but Thot did not, hence 200 translated terms, and 100 sets from 7158 that had 3 different translations each, hence 300 translated terms.

Table 5 presents the results of the analysis on the mean acceptability score of each translation, the percentage of acceptable translations, and the agreement between the two raters on this judgement.

The total mean translation score for Google Translate was 2.15. The total mean translation score of Matecat was 2.11 and of Thot 1.91. Taking 0 and 1 (regarded as not acceptable) and 2 and 3 (regarded as acceptable) together resulted in acceptability percentages between 45% and 93%.

Finally, in those cases where there was more than one translation for a term, the reviewers determined which tool they considered to provide the best translation. This is shown in Table 6.

Kappa's for acceptability were 0.623 and 0.577, meaning a fair level of agreement. Both were significant with $p<0.001$. The weighted average of the percentages from Table 5, based on the number of terms from Table 4, results in an overall percentage of acceptable terms of 61%. The translations by Matecat and Google Translate were considered better than those of Toth.

*Table 5 – Mean translation scores per reviewer, percentage of translations considered acceptable by both reviewers, kappa for acceptability and p-value of kappa.*

| Translation | Mean score (95% CI) | Acceptability Percentage acceptable | Kappa (p-value) |
|---|---|---|---|
| Equal translations | | 93% | 0.712 ($p < 0.001$) |
| - Reviewer 1 | 2.6 (2.5 - 2.8) | | |
| - Reviewer 2 | 2.7 (2.6 - 2.9) | | |
| Two different; translation Google & Matecat | | 67% | 0.400 ($p < 0.001$) |
| - Reviewer 1 | 2.1 (1.9 - 2.3) | | |
| - Reviewer 2 | 2.1 (1.9 - 2.3) | | |
| Two different; translation Thot | | 53% | 0.523 ($p < 0.001$) |
| - Reviewer 1 | 2.0 (1.8 - 2.1) | | |
| - Reviewer 2 | 1.8 (1.7 - 2.0) | | |
| All different; translation Google | | 81% | 0.277 ($p = 0.002$) |
| - Reviewer 1 | 2.4 (2.3 - 2.5) | | |
| - Reviewer 2 | 2.3 (2.1 - 2.4) | | |
| All different; translation Matecat | | 64% | 0.421 ($p < 0.001$) |
| - Reviewer 1 | 2.1 (2.0 - 2.3) | | |
| - Reviewer 2 | 2.0 (1.8 - 2.2) | | |
| All different; translation Thot | | 45% | 0.401 ($p < 0.001$) |
| - Reviewer 1 | 1.9 (1.7 - 2.1) | | |
| - Reviewer 2 | 1.7 (1.6 - 1.9) | | |

*Table 6 – Percentage of translation scored as best by both reviewers.*

| Translation | Percentage regarded best |
|---|---|
| Two different: Google & Matecat | 42% |
| Two different: Thot | 38% |
| Two different: no agreement | 20% |
| All different: Google | 39% |
| All different: Matecat | 17% |
| All different: Thot | 17% |
| All different: no agreement | 27% |

## Discussion

In this research the quality of translations by three machine translation engines was tested. To our knowledge, this is among the first studies in which different automated translations were compared.

The Thot translation improved when it was tuned with procedures. The big step to making a translation that resembled human translation was training with the available parts of Dutch SNOMED CT. This training set already contained some translations for terms that had to be translated by Thot.

Even though Thot used existing translations for SNOMED CT descriptions, it could not outperform the tools with a generic translation memory.

Translations that are the same among the three methods are generally of better quality. Ninety-three percent of those translations are acceptable, a number which is only approached by the Google translations in the case that all tools provide different translations. We expected to get less but better translations using Thot, as it was trained for this purpose. However, it has a much smaller translation memory than Google and Matecat, and the contrary has been the case. The number of translations, after filtering non-translations, was higher for Thot than for Google and Matecat, but the validity of the translations was lower.

A Java program checked the translations for non-Dutch words, and we were surprised by the large amount of terms that could be translated. This does not directly mean the translations are of good quality, but it does mean that most terms have been translated to Dutch terms.

In the comparison Matecat and Google translated many terms the same, and Thot translated very differently. This emphasizes the difference between generic and specific tools. The real quality was measured by manually assessing the validity of three samples of 100 terms. Average scores show the terms are on average considered acceptable. However, we did see that the Thot translations were the only ones to score under 2.00 on average, and under 60% in acceptable translations. This means the Thot translations were considered inferior to the other translations. There is also a difference in the total average score. The terms that were all translated the same were considered better than the other terms. The reviewers could see that these terms were shorter than other terms. This might mean there was less possibility for making different or wrong translations. When the two different translations were compared, Matecat and Google scored a bit better than Thot. When comparing the three different translations, the Google translation was considered the best most of the time. Most of the translations were considered acceptable, but not perfect. The translations made by Google, Matecat and Thot can give an idea of the translation of a term, but most of the time they will not give a translation that is ready to use. In this research, Google Translate scored the best in comparison with Matecat and Thot. Thot scored worst.

Limitations of this study were the fact that only two reviewers rated the terms, and the small size of the samples that were judged for validation of the translations. Furthermore, the analysis of acceptability of translations could be further systematized. Additionally, the types of errors in translations could be further specified. For example, the generic tools occasionally used synonyms from a non-medical domain, leading for example to translations of “blood vessels” as “blood ships.” Regarding Thot, the high inclusion rate (i.e., relatively few English terms in a translation) seems to be explained by the fact that Thot simply leaves out fragments it cannot adequately translate. This may lead to close-to-human translation, but lacks the full semantics of the English term that has to be expressed. Further analysis of such mechanisms and types of error is needed.

Creating translations with machine translation tools has the potential to help make good translations, but the translations are not made according to the translation rules for SNOMED CT from SNOMED International. This means making a translation with machine translation tools will only provide reference terms, and not official translations. This makes machine translation not suitable for making an official translation of SNOMED CT, as stated in the research of Schulz [2].

Our research contributes to the field of machine translation, in which efforts are undertaken for various languages. The distinguishing features are the application to SNOMED CT, and the translation to Dutch. In earlier research [10], we showed that Dutch is among the languages with a scarcity of resources for language processing in medicine. It similarly lacks corpora, especially bilingual medical corpora, that could contribute to improving the quality of machine translation. This poses challenges and gives need to the use of alternative approaches, such as lexical and morphosemantic approaches [11].

Our study is similar to the one described in [12], which compared three approaches for translating the Gene Ontology from English to German. They used Wikipedia, Google API with context and Google API without context. The average scores for adequacy (the extent to which a translation represents the meaning) and for fluency (the extent to which a translation is proper German) were over 4.0 on a 5-point Likert scale. This seems higher than the scores we found in our study, which may be explained by the fact that we combined adequacy and fluency in our scores, and the existence of a larger corpus of German terms, as German is the second-most represented language in PubMed/MEDLINE, after English [12].

Further research should be performed on using machine translation tools in a full translation process. This could determine whether using machine translation tools will be beneficial for translation. Training Thot using only validated translations that conform to SNOMED International’s translation guidelines may eventually lead to higher translation quality. Research could be done on building a good translation memory to train a tool such as Thot. This might result in translations of better quality. Furthermore, this may prove useful for maintenance, i.e., for providing translations for concepts that are added to new releases of SNOMED CT.

Further research is also needed to point out the impact that translations, translation quality, and adequacy of synonymy have when actually using a translation of SNOMED CT in clinical practice, for example, on impacting the inter-coder agreement, which is considered to be low when using SNOMED CT in English [13].

## Conclusion

Overall quality of the three different tools was considered acceptable, but not good enough for use in clinical practice. The Thot translations were considered worse than the Google and Matecat translations. Shorter terms were more often translated the same by the three tools, and these translations were considered better. The translations made by the tools could be used in a translation process, but cannot be used directly. The translation tools cannot translate the terms according to the translation rules for SNOMED CT. This means the tools are of limited help for making an official translation of SNOMED CT.

## References

[1] D. Lee, N. de Keizer, F. Lau, and R. Cornet, Literature review of SNOMED CT use, *J Am Med Inform Assoc* **21** (2014), e11-19.
[2] S. Schulz, J. Bernhardt-Melischnig, M. Kreuzthaler, P. Daumke, and M. Boeker, Machine vs. human translation of SNOMED CT terms, *Stud Health Technol Inform* **192** (2013), 581-584.
[3] G.A. Reynoso, A.D. March, C.M. Berra, R.P. Strobietto, M. Barani, M. Iubatti, M.P. Chiaradio, D. Serebrisky, A. Kahn, O.A. Vaccarezza, J.L. Leguiza, M. Ceitlin, D.A. Luna, F.G. Bernaldo de Quiros, M.I. Otegui, M.C. Puga, and M. Vallejos, Development of the Spanish version of the Systematized Nomenclature of Medicine: methodology and main issues, *Proc AMIA Symp* (2000), 694-698.
[4] G.O. Klein and R. Chen, Translation of SNOMED CT - strategies and description of a pilot project, *Stud Health Technol Inform* **146** (2009), 673-677.
[5] L. Deleger, T. Merabti, T. Lecrocq, M. Joubert, P. Zweigenbaum, and S. Darmoni, A twofold strategy for translating a medical terminology into French, *AMIA Annu Symp Proc* **2010** (2010), 152-156.
[6] M. Federico, N. Bertoldi, M. Cettolo, M. Negri, M. Turchi, M. Trombetti, A. Cattelan, A. Farina, D. Lupinetti, A. Martines, A. Massidda, H. Schwenk, L. Barrault, F. Blain, P. Koehn, C. Buck, and U. Germann, The Matecat Tool, *Proceedings of the 25th International Conference on Computational Linguistics (COLING)* (2014), 129-132.
[7] D. Ortiz-Martínez, and F. Casacuberta, The New THOT Toolkit for Fully-Automatic and Interactive Statistical Machine Translation, *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics* (2014), 45-48.
[8] P. Reuter, Springer Groot medisch woordenboek. Medical Dictionary. Engels-Nederlands English-Dutch. Springer Uitgeverij Houten (2008) 738 pp. ( + CD rom)
[9] K. Papineni, S. Roukos, T. Ward, and W-J. Zhu, BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th annual meeting on association for computational linguistics - ACL '02* (2002), 311-318.
[10] R. Cornet, A. Van Eldik, and N. De Keizer, Inventory of tools for Dutch clinical language processing, *Stud Health Technol Inform* **180** (2012), 245-249.
[11] O. Perez-de-Vinaspre, and M. Oronoz, Translating SNOMED CT Terminology into a Minor Language, *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis, Louhi (2014)*: 38-45.
[12] N.D. Hailu, K.B. Cohen, and L.E. Hunter, Ontology translation: a case study on translating the gene ontology from English to German, *International Conference on Applications of Natural Language to Data Bases/Information Systems* (2014), 33-38.
[13] J.E. Andrews, R.L. Richesson, and J. Krischer, Variation of SNOMED CT coding of clinical research concepts among coding experts, *J Am Med Inform Assoc* **14** (2007), 497-506.

### Address for correspondence

Ronald Cornet
Academic Medical Center – University of Amsterdam
Department of Medical Informatics
Amsterdam Public Health research institute
P.O. Box 22700
1100 DE Amsterdam
The Netherlands
r.cornet@amc.uva.nl