This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402.

D5.6: Report on Third Year's Evaluation

Author(s): Alexandra Birch, Juliane Ried, Colin Davenport, Matthias Huck, David Mareček

Dissemination Level: Public

Date: January, 31st 2018

Version 1.0

Grant agreement no.		644402
Project acronym		HimL
Project full title		Health in my Language
Funding Scheme		Innovation Action
Coordinator		Barry Haddow (UEDIN)
Start date, duration		1 February 2015, 36 months
Distribution		Public
Contractual date of delivery		January, 31st 2018
Actual date of delivery		January, 31st 2018
Deliverable number		D5.6
Deliverable title		Report on Third Year's Evaluation
Type		Report
Status and version		1.0
Number of pages		43
Contributing partners		UEDIN, NHS 24, Cochrane, Lingea, CUNI
WP leader		UEDIN
Task leader		UEDIN
Authors		Alexandra Birch, Juliane Ried, Colin Davenport, Matthias Huck, David Mareček
EC project officer		Tünde Turbucz
The Partners in HimL are:	The University of Edinburgh (UEDIN), United Kingdom	
	Univerzita Karlova V Praze (CUNI), Czech Republic	
	Ludwig-Maximilians-Universitaet Muenchen (LMU-MUENCHEN), Germany	
	Lingea SRO (LINGEA), Czech Republic	
	NHS 24 (Scotland) (NHS24), United Kingdom	
	Cochrane (COCHRANE), United Kingdom	

For copies or reports, updates on project activities and other HimL-related information, contact:

Barry Haddow
University of Edinburgh

bhaddow@staffmail.ed.ac.uk
Phone: +44 (0) 131 651 3173

Contents

1 Introduction

The aim of this report is to describe the evaluation work performed on systems that have been investigated and deployed during the final year of the project.

In the third year of the project we have run an extensive range of human evaluations. In order to plan the costly human evaluations we developed a Year 3 evaluation plan. We show the Gantt chart we used to co-ordinate the evaluation effort in Figure 1. The different sections of this deliverable follow the individual evaluations shown in the chart. The first evaluation, ranking, is included in the first section of the report, Section 2, which measures the overall accuracy of translation models in the HimL project. In this section we compare the output of research systems, deployment systems, and commercial systems. For all the subsequent sections in this report we evaluate the output of the official HimL Year 3 system deployed by Lingea.

We start in Section 2 by presenting the overall results of the HimL project. We focus on a human ranking evaluation of the quality of translated sentences from translation systems, which are interesting either for their research contribution, or because they have been integrated into the HimL platform and deployed by users. We also include evaluation of a commercial translation platform, Google. We perform further automatic analysis of translation systems using standard metrics BLEU and Chrf, and the automatic semantic metrics AutoDA and TreeAggreg. Both human and automatic metrics show that our deployed systems have improved dramatically over the time period of the project, and that we are competitive with, or better than commercial translation systems for all four target languages.

In the Cochrane Post-Editing Evaluation Section, Section 3, we show that Cochrane's professional translators are more efficient translating from machine translations, than when translating from scratch. In the Cochrane User Survey section, Section 4, we aim to determine whether MT output itself would be useful to Cochrane users. Results show that to 75% of German respondents and about 50% of Czech and Romanian respondents prefer the HimL machine translations over English text only, suggesting that a substantial number of users would benefit from the provision of machine translated Cochrane content, especially for users with lower levels of English proficiency. Polish acceptance rates were lower as the quality of Polish translations seems lower than the other languages.

The NHS24 User Survey section, Section 5, describes an evaluation which was run to determine how people access and use health information and NHS services, and also asks them to evaluate the usefulness of machine translations of NHS24 web pages. In the NHS24 Needs Impact Study, Section 6, we describe in-depth interviews which were conducted in order to determine what information needs NHS24 Polish and Romanian users have, and if and how machine translation should be incorporated into NHS24 services. Results show that users expect completely accurate information on NHS24 branded websites and that currently we would need to employ humans to post-edit MT before publication.

The final content section, Section 7, reports statistics from NHS24 website and the Cochrane website, focusing on the countries from which people are accessing the information, and the browser language.

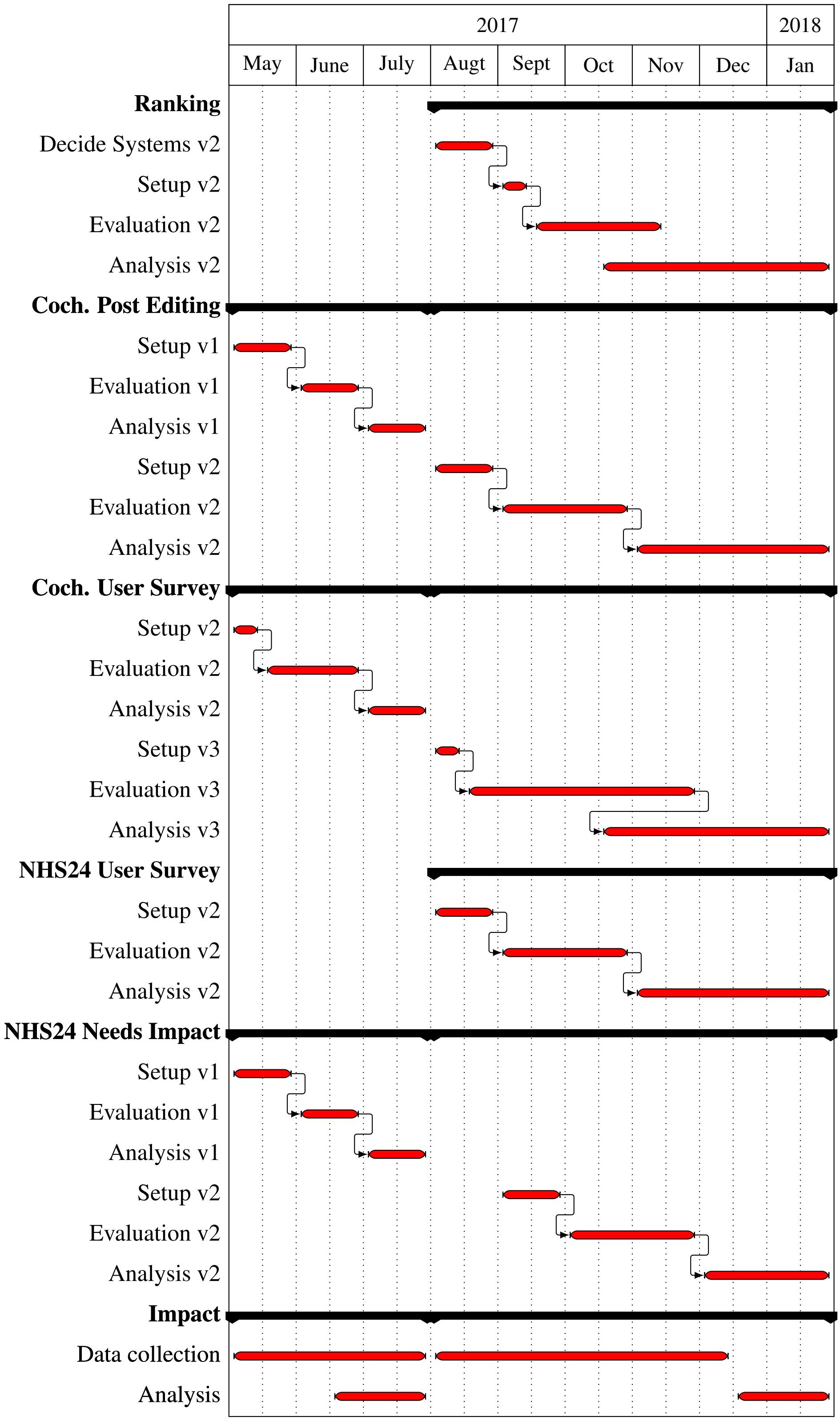


Figure 1: Schedule for HimL Evaluations

2 Accuracy Evaluations

Human evaluation of different HimL systems is essential to measure the real impact of our research on the accuracy of the translations. In this section we report the human ranking results of the HimL project and compare this to automatic evaluation results. We compare the deployed systems with other systems developed as part of research related to HimL. We additionally compare our systems to Google translations.

This evaluation is very similar to the evaluation run at the end of Year 2 of the project which was reported on in Deliverable D5.4, but we use the new test sets and we add more research systems into the evaluation. In total we rank about four times more translations than last year. We also add ChrF evaluation and semantic automatic evaluations. Any details which have not changed since last year are not described here and we ask the reader to refer to D5.4.

2.1 Human Ranking Y3

2.1.1 Data

We used both the HimL 2015 and the 2017 test sets to run these experiments. The entire test set was converted into xml HITs (Human Intelligence Task) which consist of a ranking task for 3 consecutive sentences. These HITs were then randomly shuffled so that annotators could not guess which system they were annotating by its position in the list. In Table 1 we can see the number of sentences and HITs created for the ranking task.

	Cochrane	**NHS24**
Sentences	2274	4602
HITs	758	1534

Table 1: Number of sentences and HITs that were used in the ranking task

2.1.2 Annotators

NHS24 ran a second community engagement process to recruit evaluators for Polish and Romanian translators, which resulted in somewhat fewer annotators than last year. Invitations to take part in the ranking task were distributed to all previous annotators and support agencies, with new engagement coming from social media and suggested contacts from the wider Partnership and Engagment team at NHS 24, Scottish Health Council and NHS 24 Participation and Equalties Manager. Some recruitment issues were encountered this year, with previous annotators being unable to take part through illness. As of 30th November 2017, the task had been accessed by five Polish annotators and one Romanian annotator.

The Romanian annotators were contacted via sources from organisations including Govanhill Housing Association and Crossreach. The Romanian community is not an easy one to engage with, and has well-known issues with poverty and illiteracy. As such, after initial contact from NHS 24 there is often little or no response. Despite repeated prompting only one Romanian annotator took part in the Y3 ranking task. The Polish annotators came from various sources such as Universities, Health and Social Care and Service Users (via support organisations such as the Polish Family Support Centre (PFSC)). As with previous years, the Polish community is easier to engage with due to greater integration in Scottish society. Many Polish people contacted were not interested in taking part in the project. One of the reasons given was a preference to return to Poland for treatment instead of using the NHS in Scotland. As with Y2, the selection process for the volunteer annotators consisted of a short interview to confirm instructions and expectations for the task were understood. They annotators were then given their unique token for the Appraise system along with instructions on how to complete the task. NHS 24 continued to utilise High Street 'Love2Shop' Vouchers to incentivise participation. After taking part in tasks, each annotator will be given a voucher (£10 for each task completed).

Cochrane provided professional translators and medical experts who were contracted to perform 40 hours of evaluation for all of the target languages.

In Table 2 we can see the number of sentences ranked by different annotators. A sentence rank contains ranks for 5 system translations.

2.1.3 Systems

Here we describe the translation models which we evaluate using human ranking judgments. These systems have been selected either because they were important for the project (ie. Year 2 HimL deployed system) or because they highlighted some

Target Language	Annotator Id	Sentences Ranked
German	1	738
	2	477
Total		1215
Czech	1	1083
	2	375
Total		1458
Polish	1	615
	2	594
	3	278
	4	150
	5	36
	6	33
	7	10
	8	6
Total		1722
Romanian	1	723
	2	600
	3	45
Total		1368

Table 2: Number of sentences that were ranked per annotator. Each sentence ranked 5 different translations.

interesting research performed in the project.

- **Year 2:** German, Czech, Polish, Romanian
  Phrase-based MT system, described in D4.2/4.5.
- **Year 3:** German
  The deployed HimL Year 3 engine—as described in detail in D4.3/4.6—is a neural MT system with one major novel characteristic: a linguistically informed word segmentation technique is applied to the German target language side of the training data (Huck *et al.*, 2017a,b). The word segmentation technique cascades a suffix splitter, a compound splitter, and BPE. The Year 3 NMT model is first trained on large general-domain data and then fine-tuned towards the medical domain, using a combination of the in-domain sections of the corpus and a synthetic data mix. As a last step, the *n*-best output of the left-to-right model is reranked with a right-to-left model.
- **Year 3:** Czech, Polish, Romanian
  Neural MT systems, described in D4.3/4.6. For Polish and Romanian this is a deep NMT model, and for Czech it is a shallow model.
- **Google:** German, Czech, Polish, Romanian
  The Google translations were created on the 15 September 2017. We split the files into allowable chunks and then translated the files via the API. At this point Google had converted all of these models to neural systems. Unfortunately using a commercial system as a baseline is not very informative as the exact nature of the data and model is unknown, and they are constantly changing their products.
- **Year 3 Baseline:** German, Polish, Romanian
  The Year 3 Baseline was the NMT system that we used to test the initial deployment. It is a "standard" (i.e. shallow) Nematus system, trained on a large parallel corpus extracted from OPUS and WMT, and fine-tuned with back-translated synthetic data selected from CommonCrawl. The systems are more fully described in D1.1, and we selected the best-performing system from Table 11.
- **Transformer:** German, Czech, Polish, Romanian
  The Transformer model (Vaswani *et al.*, 2017) implemented in the Tensor2tensor framework[1] is described in D2.3.
- **MLFix:** German, Czech
  The statistical automatic post-editing component, described in D3.1/3.3. The general architecture of the component is based on its predecessor, Depfix. The component was applied to the outputs of the Year 3 system for German and WMT17 system for Czech.

[1] https://github.com/tensorflow/tensor2tensor

- **Reconstruction:** Polish, Romanian
  The reconstruction system is a variant of the Year 3 neural MT system. The distinguishing feature is that it adds a 'reconstructor' (Tu *et al.*, 2016) component (similar to an auto-encoder) to the standard attentional encoder-decoder architecture. The aim is to improve the adequacy of translations by encouraging the model to extract more information from the source sentence in order to avoid the common problem of under-translation, where parts of the source sentence are left untranslated. The model is trained via continued training of a non-reconstruction model with a modified objective function that rewards accurate source sentence reconstruction. During decoding, reconstruction scores are used to rerank the initial *n*-best output.

- **Inverse:** Polish, Romanian
  The inverse system has a similar motivation to the reconstruction system, but uses a standard attentional encoder-decoder architecture. Two models are trained, one in the desired 'forward' direction (e.g. English to Polish) and one in the inverse direction (Polish to English). During decoding, the inverse system is used to rerank the *n*-best output of the forward system.

- **WMT17:** Polish, Czech, Romanian
  For Romanian and Polish this is essentially the same as the y3 system, except we perform an additional right-to-left reranking step, i.e. we rescore the output with an ensemble of systems trained to produce output right-to-left, and rerank. For Czech, this is not actually our WMT17 submission – it is a new system built in the same way as the Polish and Romanian systems WMT17, which improves on the y3 system by using a deep model. The systems are described in (Sennrich *et al.*, 2017).

- **Depfix:** Czech
  A rule-based automatic post-editing component described in Rosa (2014) focused on fixing English to Czech SMT errors. Depfix, being the predecessor of the MLFix, is using a set of hand-crafted rules to fix morphology, negation and other errors instead of statistical components. Depfix was applied on the outputs of the WMT17 System.

- **Form2Form:** Czech
  This is a baseline system for experiments using semantic role labelling (e.g. the system Intvfid), described in detail in deliverable D2.3.

- **Intvfid:** Czech
  Neural MT system using interleaved valency frames of verbs in the source sentences, comparable with the previous Form2Form baseline. It is described in detail in deliverable D2.3.

2.1.4 Method

The ranking evaluation was performed in order to obtain human judgements comparing alternative HimL translation systems.

We follow (Bojar *et al.*, 2016) to extract ranking results from the raw ranking data of five ranked system for each test sentence. From these rankings, we produce pairwise translation comparisons,

While the precise method of combining rankings over different systems and sentences has varied over the years, it has always shared a common idea of computing the average number of times each system was judged better than other systems, and ranking from highest to lowest. Introduced for WMT13 (Macháček and Bojar, 2013), the Expected Wins (EW) has an intuitive score demonstrated to be accurate in ranking systems according to an underlying model of relative ability. The idea is to gauge the probability that a system S_i will be ranked better than another system randomly chosen from a pool of opponents $S_j : j \neq i$. If we define the function $win(A, B)$ as the number of times system A is ranked better than system B then we can define this as follows:

$$EW(S_i) = \frac{1}{|S_j|} \sum_{j,j\neq i} \frac{win(S_i, S_j)}{win(S_i, S_j) + win(S_j, S_i)} \tag{1}$$

Note that this ranking aggregation approach disregards ties. In the year 2 evaluation report (D5.4) we reported TrueSkill results. TrueSkill uses the relative strength of two systems when updating their scores, but this method is not as intuitive as expected wins, and led to some unstable results.

We perform an analysis of the performances of these systems as reported by the industry standard BLEU score. We used the mteval13b script in the MOSES software to compile these numbers.

We used the ChrF metric which is a character n-gram F-score which has shown the highest segment-level correlation with human judgements on the WMT14 shared evaluation task. It works especially well for translations out of English and into languages with rich morphology, making it especially suitable as an auxiliary metric for HimL. We took the evaluation script from the sub-word repository [2].

All evaluation (human and automatic) is applied to detokenised and recased output.

2.1.5 Results

English-German

System	Ranking	BLEU	ChrF
Year 3	**0.662**	36.65	64.02
MLFix	0.629	36.64	64.02
Transformer	0.591	37.02	64.13
Year 3 Base	0.537	35.04	62.48
Google	0.425	**37.57**	**64.60**
Year 2	0.154	28.85	60.07

English-Czech

System	Ranking	BLEU	ChrF
Transformer	**0.702**	**31.69**	**57.96**
WMT17	0.640	29.83	56.49
MLFix	0.616	29.90	56.48
Google	0.565	29.97	56.55
DepFix	0.547	29.78	56.45
Form2Form	0.443	26.59	53.69
Year 3	0.431	27.88	54.71
IntvFid	0.400	27.00	54.24
Year 2	0.154	20.86	49.70

English-Polish

System	Ranking	BLEU	ChrF
Reconstruction	**0.653**	**25.11**	**52.98**
WMT17	**0.653**	25.03	52.65
Year 3	0.572	24.98	52.78
Inverse	0.635	24.85	52.95
Transformer	0.570	23.07	51.23
Google	0.434	22.89	51.02
Year 3 Baseline	0.339	22.14	50.68
Year 2	0.144	19.11	49.06

English-Romanian

System	Ranking	BLEU	ChrF
Transformer	**0.684**	30.24	59.38
Google	0.597	**38.40**	**64.23**
Reconstruction	0.572	36.33	61.40
Inverse	0.562	36.58	61.82
Year 3	0.489	35.43	60.35
WMT17 Corrected	0.493	36.00	60.65
Year 3 Baseline	0.319	32.24	58.64
Year 2	0.284	32.36	60.86

Table 3: Expected wins ranking results on a concatenation of 2015 and 2017 test sets for both Cochrane and NHS24. We have included BLEU and ChrF scores for the same test set in these tables in order to compare the human and automatic evaluation results.

In Table 3 we can see the results of the ranking experiment and of the automatic BLEU and ChrF scores on the entire HimL test set, containing both the 2015 and 2017 test sets and the Cochrane and NHS24 portions. We describe the most important results:

- We have significantly improved over the Year 2 systems for all four target languages, and for both ranking and automatic scores. This demonstrates our considerable progress in the second half of the HimL project.
- HimL systems beat Google for all of the languages, even though Google performs particularly well for German and Romanian, coming first in the automatic metric ranking but not in the human ranking.
- Generally speaking there is a strong correlation between the performance of systems as measured by humans and by the automatic scores. There are some notable exceptions:
  - Google's German system has high BLEU scores but performs poorly on human ranking. It could be that the Google system is not well adapted to the HimL domain and the domain adaptation techniques applied in our Year 3 system are not being rewarded strongly enough by the BLEU metric.
  - The second divergence is the Romanian Transformer system. Here the BLEU score is quite poor but it still ranks first for human evaluation. This could be due to the problem of diacritics and the Transformer model not training on data consistent with the test set. These diacritics are not likely to be penalised by humans as harshly as the automatic metrics.
  - The final important divergence relates to the Reconstruction model. This model performs very well in human rankings, and not as well on automatic rankings. This is potentially a very interesting result because the Reconstruction model specifically uses extra information from the source sentence to reward translations, and indicates that we are potentially reducing the problem of leaving difficult parts of the sentence untranslated.

[2] https://github.com/rsennrich/subword-nmt/blob/master/chrF.py

- Unlike in the Year 2 evaluation, this year there are many strong models competing and in general there is a big group of systems with similar performance at the top of each table. The phrase-based Year 2 system is significantly worse, and also the Year 3 baseline system is one of the weaker systems.
- The deployed Year 3 system is a strong system, but only ranks first in German. For the other languages research systems, and sometimes Google, beat the deployed system. There are a number of reasons for this:
  - Some research systems were developed after the Year 3 system was deployed (like the Transformer model)
  - Some research systems were not selected for deployment due to their complicated requirements or decoding speed (like the WMT17 system).
  - The final important reason that the Year 3 system is sometimes not the top scoring system is that we had to rely on automatic scores to choose the system to deploy and some systems are favoured more by humans than by automatic scores.
- For **English-German**, we have achieved a massive improvement over Year 2. The gain in translation quality is very evident both in automatic scores and in terms of human judgement. According to human judgement, the deployed HimL English-German Year 3 system is also ranked higher than Transformer and Google, despite a slightly lower BLEU score. We conjecture that the HimL Year 3 system benefits from its linguistically informed target word segmentation. A similar effect was observed in the WMT17 news translation shared task (Bojar *et al.*, 2017).
- In **English-Czech**, the Y3 system is lower quality since it is a shallow NMT system, whereas the WMT17 version is a deep system.
- In **English-Romanian**, our system suffered from noisy training data, caused ultimately by inconsistent orthography in Romanian. We made some effort to correct this, but there are still residual problems with the system.
- The difference between the Y3 and WMT17 systems for **English-Polish** and **English-Romanian** is just that the WMT17 systems have an extra step where *n*-best output is reranked by another NMT system that produces reversed target language. This reranking step helps to mitigate against cases where the NMT system chooses to ignore parts of the source sentence, or to over-generate.

We developed the HimL 2017 test set in order to test systems on novel test data, as some of the systems tested have used the HimL 2015 test set for tuning and development. We also concentrate on the test set for which we have a good number of judgements across the four languages, the Cochrane test set. In Table 4 we can see these results.

In general the results in this table agree with the results in the previous table, Table 3, which confirms the conclusions we have drawn from the previous results. The most notable difference is that the BLEU scores are higher for the Cochrane test set which might be due to the fact that Cochrane text is more similar to standard training sets. The NHS24 data has lists and titles and is more removed from the standard written text training sets.

2.2 Three Years of HimL Models

Since this is the final report on evaluation of the HimL translations, we summarise the work that we have done in the project by reporting here on the results of all the deployed systems over the last three years and compare them to the two Google systems we have used. We only use the HimL 2015 test set as we have Y1 and Google 2016 results for this test set.

In Table 5 we can see the scores assigned to the different systems across the HimL 2015 test set. We have selected a subset of the systems to make it easy to compare progress over the entire length of the project across all languages. We can see that we have made dramatic progress. The BLEU score improvement between Y1 and Y3 systems is more than 4 points, with the greatest improvement seen in Romanian where we have a 12 point improvement.

The commercial systems are also constantly improving. Google 2017 system improves up to 6 points over Google 2016 system. Although Google does have a BLEU score advantage for Czech and Romanian, we beat Google in the human ranking evaluation in our deployed Y3 system for two language pairs, and across the board when considering our latest research systems (see Table 3).

In order to understand these results better we have broken them down by looking at BLEU scores for the Cochrane and NHS24 test sets separately. In Table 6 we can see results for the Cochrane sentences and in Table 7 we can see results for the NHS24 sentences.

For the Cochrane results we can see that Google beats us on three of the 4 language pairs, as measured by BLEU. Even with our adaptation of the systems to the public health domain, Google's access to large amounts of training data and almost unlimited computing resources help them to perform really well on automatic metrics. However, human evaluation prefers our Y3 system for both Polish and German (see Table 4), and we do even better with our recent research systems.

English-German

System	Ranking	BLEU	ChrF
Year 3	**0.689**	39.98	67.24
MLFix	0.609	39.98	67.24
Transformer	0.592	40.45	68.27
Year 3 Base	0.534	36.53	65.08
Google	0.422	**44.29**	**69.98**
Year 2	0.154	29.17	61.13

English-Czech

System	Ranking	BLEU	ChrF
Transformer	**0.707**	**32.65**	**60.86**
MLFix	0.672	31.45	60.24
WMT17	0.597	31.45	60.24
DepFix	0.569	31.37	60.24
Google	0.534	31.89	60.08
Form2Form	0.457	26.92	56.93
Year 3	0.413	28.08	57.23
IntvFid	0.397	28.22	57.21
Year 2	0.155	19.90	50.85

English-Polish

System	Ranking	BLEU	ChrF
Reconstruction	**0.685**	**28.62**	**57.67**
WMT17	0.671	28.58	57.47
Inverse	0.626	28.58	57.53
Transformer	0.591	25.81	55.54
Year 3	0.551	27.24	56.79
Google	0.412	24.09	53.72
Year 3 Baseline	0.331	24.82	54.36
Year 2	0.132	18.09	51.07

English-Romanian

System	Ranking	BLEU	ChrF
Transformer	**0.688**	36.63	65.53
Reconstruction	0.598	42.24	66.44
Google	0.581	**45.21**	**69.39**
Inverse	0.531	42.43	66.88
WMT17 Corrected	0.521	41.71	66.04
Year 3	0.475	40.89	65.68
Year 3 Baseline	0.308	36.88	62.88
Year 2	0.296	32.46	63.16

Table 4: Expected wins ranking results on the final 2017 test set for Cochrane which contains more ranking judgments. Systems are ordered by their inferred system means. We have included BLEU and ChrF scores (over the 2017 Cochrane test sets) in these tables in order to compare the human and automatic evaluation results.

System	German	Czech	Polish	Romanian
Y1	32.07	22.07	19.45	26.86
Y2	30.95	23.49	21.23	34.93
Y3	**36.52**	30.06	**25.28**	36.75
Google 2016	35.36	27.06	21.29	32.75
Google 2017	36.24	**31.89**	23.60	**38.16**

Table 5: BLEU scores results for the HimL evaluation on the HimL 2015 test data.

For the NHS24 results, we beat Google on two of the 4 language pairs. Here our Czech and Romanian systems are beaten by Google, but our Polish and German Y3 system is stronger than all the rest. In fact the Y3 system seems to cope very well with the NHS24 data, coming first for Polish and Romanian.

2.3 Semantic Automatic Evaluation

We created and evaluated two new automatic MT metrics: **AutoDA** and **TreeAggreg**.

AutoDA is a sentence-level metric trainable on any direct assessment scores. The metric is based on a simple linear regression combining several features extracted from the automatically aligned translation-reference pair. There may be also other established metrics within the features. AutoDA was described in detail already in previous Deliverable 5.3 in Section 3. There are two variants of the AutoDA metric. One is language universal and requires a treebank existing in the Universal Dependencies collection (Nivre *et al.*, 2016). It is applicable to all four HimL languages. The language universal features are given in D5.3, Section 3.3. The other version is called **AutoDA.tecto** and incorporates deeper syntactic features from the tectogrammatical layer of annotation, which is available only for Czech. Therefore it cannot be used for the other three languages. It was described in D5.3, Section 3.2. Both AutoDA metrics were trained on WMT Direct Assessment scores (Bojar *et al.*, 2016) or HUMEseg scores and participated in the WMT17 Metrics task.

TreeAggreg is a simple sentence-level metric, inspired by HUME (Birch *et al.*, 2016). Rather than being a full standalone metric, it can be regarded as a *metric template*, for in principle, any string-based MT metric can be plugged into it; we used chrF3 (Popovic, 2015). In TreeAggreg, we are trying to improve an existing string-based metric by applying it in a syntax-tree-

System	German	Czech	Polish	Romanian
Y1	33.82	23.68	15.72	27.78
Y2	35.55	25.55	17.01	37.09
Y3	39.25	33.48	**22.52**	39.08
Google 2016	37.56	29.78	18.47	35.39
Google 2017	**39.47**	**35.68**	20.82	**40.01**

Table 6: BLEU score results for the HimL Year 3 evaluation on Cochrane test data

System	German	Czech	Polish	Romanian
Y1	30.09	20.42	23.58	25.87
Y2	26.23	21.34	25.37	32.55
Y3	**33.74**	26.44	**28.34**	34.09
Google 2016	33.10	24.17	24.37	29.27
Google 2017	32.92	**27.87**	26.68	**35.47**

Table 7: BLEU score results for the HimL Year 3 evaluation on NHS24 test data

based context. This is motivated by our belief that dependency trees are a good means of capturing sentence structure, which may be relevant for MT evaluation metrics, as the MT output should presumably transfer the information present in the source sentence into a similar syntactic structure as the reference translation uses. However, in string-based MT metrics, the syntactic structure of a sentence is typically ignored. In our rather light-weight attempt to employ syntactic analysis in MT evaluation, we segment the sentences into phrases based on their dependency parse trees, and evaluate these phrases independently with the string-based MT metric. The resulting scores are then aggregated into a final sentence-level score using a simple weighted average. TreeAggreg also participated in the WMT17 Metrics task.

A detailed description and the results of these two metrics are available in the enclosed paper (Mareček *et al.*, 2017)

The results of AutoDA and TreeAggreg metrics are compared to the rankings in Table 8. We see that the semantic metrics generally follow the results given by the BLEU scores, even in cases where the human ranking results are different to the BLEU scores results (See English-German). There is a slight discrepancy with BLEU in English-Polish, where both the semantic metrics prefer the Inverse model over WMT17 and the Reconstruction model for the entire test set. It is reassuring to have access to automatic metrics which focus on different aspects of translation quality and to be able to report scores across a number of different metrics.

2.4 Discussion

We have reported on comprehensive human accuracy evaluations as shown by human ranking, and automatic evaluations and complemented this with evaluations using our semantic automatic metrics.

The results from this section indicate that enormous progress has been made over the three years of this project. We have shown that our deployed systems have improved by up to 12 BLEU points from the strong Y1 systems to the Y3 systems. Our human ranking results show that our latest research systems consistently beat previous years' deployed systems and a strong commercial system, even without access to the training data and compute resources available to Google.

The most interesting models have shown to be the non-recurrent Transformer model and the semantically-inspired Reconstruction models and these models show that we cannot rely purely on BLEU score evaluation, as these models are penalised by the automatic metrics.

English-German

	Cochrane 2017				Cochrane+NHS24 2015+2017			
System	Ranking	BLEU	AutoDA	TreeAggreg	Ranking	BLEU	AutoDA	TreeAggreg
Year 3	**0.689**	39.98	0.7704	0.6563	**0.662**	36.65	0.7451	0.6310
MLFix	0.609	39.98	0.7704	0.6563	0.629	36.64	0.7501	0.6310
Transformer	0.592	40.45	0.7788	0.6725	0.591	37.02	0.7472	0.6265
Year 3 Base	0.534	36.53	0.7600	0.6386	0.537	35.04	0.7401	0.6121
Google	0.422	**44.29**	**0.7877**	**0.6876**	0.425	**37.57**	**0.7531**	**0.6361**
Year 2	0.154	29.17	0.7400	0.5976	0.154	28.85	0.7241	0.5800

English-Czech

	Cochrane 2017				Cochrane+NHS24 2015+2017			
System	Ranking	BLEU	AutoDA	TreeAggreg	Ranking	BLEU	AutoDA	TreeAggreg
Transformer	**0.707**	**32.65**	**0.8025**	**0.6038**	**0.702**	**31.69**	**0.7687**	**0.5605**
MLFix	0.672	31.45	0.8000	0.5963	0.616	29.90	0.7675	0.5579
WMT17	0.597	31.45	0.8001	0.5963	0.640	29.83	0.7676	0.5578
DepFix	0.569	31.37	0.8002	0.5963	0.547	29.78	0.7674	0.5575
Google	0.534	31.89	0.7967	0.5905	0.565	29.97	0.7703	0.5572
Form2Form	0.457	26.92	0.7866	0.5671	0.443	26.59	0.7540	0.5287
Year 3	0.413	28.08	0.7867	0.5662	0.431	27.88	0.7586	0.5364
IntvFid	0.397	28.22	0.7865	0.5680	0.400	27.00	0.7542	0.5290
Year 2	0.155	19.90	0.7511	0.5004	0.154	20.86	0.7361	0.4969

English-Polish

	Cochrane 2017				Cochrane+NHS24 2015+2017			
System	Ranking	BLEU	AutoDA	TreeAggreg	Ranking	BLEU	AutoDA	TreeAggreg
Reconstruction	**0.685**	**28.62**	**0.6575**	**0.5756**	0.544	**25.11**	0.6298	0.5293
WMT17	0.671	28.58	0.6571	0.5746	**0.653**	25.03	0.6304	0.5299
Inverse	0.626	28.58	0.6571	0.5746	0.635	24.85	**0.6308**	**0.5320**
Transformer	0.591	25.81	0.6470	0.5560	0.570	23.07	0.6170	0.5072
Year 3	0.551	27.24	0.6526	0.5684	0.572	24.98	0.6226	0.5313
Google	0.412	24.09	0.6396	0.5324	0.434	22.89	0.6201	0.5114
Year 3 Baseline	0.331	24.82	0.6406	0.5390	0.339	22.14	0.6150	0.4990
Year 2	0.132	18.09	0.6164	0.4985	0.144	19.11	0.5984	0.4780

English-Romanian

	Cochrane 2017				Cochrane+NHS24 2015+2017			
System	Ranking	BLEU	AutoDA	TreeAggreg	Ranking	BLEU	AutoDA	TreeAggreg
Transformer	**0.688**	36.63	0.7936	0.6385	**0.684**	30.24	0.7386	0.5649
Reconstruction	0.598	42.24	0.8013	0.6560	0.117	36.33	0.7555	0.5950
Google	0.581	**45.21**	**0.8144**	**0.6799**	0.597	**38.40**	**0.7690**	**0.6188**
Inverse	0.531	42.43	0.8022	0.6560	0.562	36.58	0.7596	0.6019
WMT17 Corrected	0.521	41.71	0.7985	0.6509	0.493	36.00	0.7525	0.5900
Year 3	0.475	40.89	0.7977	0.6508	0.489	35.43	0.7461	0.5856
Year 3 Baseline	0.308	36.88	0.7790	0.6165	0.319	32.24	0.7415	0.5698
Year 2	0.296	32.46	0.7717	0.6114	0.284	32.36	0.7428	0.5785

Table 8: Results of the automatic MT metrics AutoDA and TreeAggreg, compared to the ranking scores and BLEU scores (see Tables 3 and 4). The first four columns show the results on the Cochrane 2017 dataset. The other four columns show results on a concatenation of all the test sets from 2015 and 2017, Cochrane and NHS24. Systems are ordered according to the rankings on the Cochrane 2017 dataset.

3 Cochrane post-editing evaluation

Following the post-editing pilot evaluation conducted in June 2017, which was reported on in Deliverable 5.5, Cochrane repeated and expanded the experiment for the final year 3 evaluation. The aim was to confirm and improve on the promising results from the pilot with a larger dataset, optimized HimL MT engines, and improved data collection, to demonstrate that post-editing HimL MT is less effort and quicker than Cochrane's standard translation workflow. This would allow Cochrane to publish more translations of its health information faster in the HimL languages, and reduce resources needed for its translation activities.

3.1 Experiment design and setup

The experiment design and setup was largely the same as described in D5.5 for the pilot experiment. However, Cochrane used the final HimL Y3 systems as the basis for the post-editing task, and expanded the experiment to include ten Cochrane Plain Language Summaries (PLSs), compared to only three PLSs in the pilot. Each of these ten PLSs were translated twice into the four HimL languages by two different translators respectively: once by post-editing HimL MT, and once by translating from scratch, but with access to Google MT and, for Polish and German, Cochrane's existing TM and glossaries. This approach allowed Cochrane to compare post-editing effort and time taken to edit HimL MT with Cochrane's standard translation workflow.

Translators were instructed to distribute the tasks for their language according to their preference, but to ensure that they would not be working on the same PLS twice, i.e. that no translator would complete the post-editing and standard workflow task for the same PLS. Additionally, they were instructed on best practice in how to use MateCat to avoid incorrect time-to-edit recordings, a problem that had occurred in the post-editing pilot and somewhat skewed the collected data. The translators had a little over two months to complete the post-editing task.

3.2 Results from the post-editing task

The editing log of each translation was exported from MateCat in full. As before, the analysis focused on time-to-edit, average number of seconds spent on editing per word, and post-editing effort (PEE), which was available in the editing log per segment, i.e. per sentence or header. MateCat defines PEE as the overall percentage of the pre-translated content (taken from either a TM or the MT engine) amended by the translator[3]. The totals and averages for those data were calculated by task type and language and are presented in Table 9. A full break-down of results by language and task is available in Appendix A.1. The PEE for the standard workflow tasks has been omitted from the results, as it isn't applicable: A translation "from scratch" doesn't include pre-translated content.

Post-editing was clearly quicker than standard human translation for Czech, German, and Romanian. For Czech, post-editing reduced the required translation time by about 30% on average compared to standard human translation (7 average seconds per word versus 10 average seconds per word). This result was similar in the post-editing pilot, in which a 34% reduction in editing time was recorded for Czech. For German, post-editing reduced the required translation time by about 40% on average compared to standard human translation (3 average seconds per word versus 5 average seconds per word), which was a positive result, but lower than the 57% recorded in the post-editing pilot. For Romanian, the time reduction amounted to 60%, compared to 33% in the pilot (2 average seconds per word versus 5 average seconds per word). For Polish, post-editing and standard human translation took about the same time, which was already the case in the pilot project, so the second experiment could not improve on the earlier results for Polish.

The average PEE in the German and Romanian post-editing tasks decreased from 21% to 12%, and from 18% to 11% respectively compared to the post-editing pilot. This suggests that the HimL Y3 systems produced more accurate German and Romanian translations than the neural Y3base systems that were used in the pilot, assuming the post-editors worked to the same standards in both experiments. On the other hand, the average PEE increased for Czech and Polish, from 21% to 28%, and from 33% to 35% respectively compared to the pilot project.

3.2.1 Translator feedback survey

Upon completion of the post-editing task, Cochrane invited translators to provide feedback via a brief qualitative survey in SurveyMonkey. These results are presented in Table 10. Ten out of eleven translators provided feedback. All respondents, with the exception of one of the Polish translators, rated HimL machine translations for the purpose of post-editing as acceptable and/or helpful. When asked which approach they thought had been quicker out of translating from scratch or post-editing, only one Polish translator selected translating from scratch, while seven out of ten translators thought post-editing HimL machine translation was faster. The other two translators were not sure, or said neither approach was faster. When given the chance to add any final comments, the German translators in particular were very positive about the machine translation, noting it was

[3] see https://www.matecat.com/support/advanced-features/editing-log/ (last accessed on 10 January 2018)

	Czech	German	Polish	Romanian
Post-editing				
Total Words	4871	5293	5328	5336
Total time-to-edit (hh:mm:ss)	10:27:19	03:39:56	07:05:09	03:05:30
Avg secs/word	7	3	5	2
Avg PEE	28%	12%	35%	11%
Standard Workflow				
Total Words	4115	5305	4891	2628
Total time-to-edit (hh:mm:ss)	10:59:58	07:57:27	05:58:46	03:10:00
Avg secs/word	10	5	5	5

Table 9: Summary of results from post-editing pilot by language and task type. PEE = Post-editing effort.

"considerably easier and quicker" compared to translating from scratch, while another commented that the quality of HimL machine translation was superior to the Google Translate option available in the translation management system usually used by Cochrane translators. One of the Czech translators described the post-editing as "very efficient and time-saving". Echoing the results from the editing log presented above, there was a comment from one of the Polish translators that the machine translations did not work very well for Polish.

Despite the largely positive feedback on quality of as well as speed-up from HimL MT, three of the ten translators said they preferred translation from scratch over post-editing, and that included Czech, German and Polish respondents.

3.2.2 Limitations of the collected data

The best practice instructions given to translators to avoid incorrect time-to-edit records helped improve the data, but there were still some segments with average editing times of over 25 seconds per word. As those segments most likely represent instances where translators interrupted their work, those were excluded from the data.

While the evaluation was still in progress, one of the Czech translators reported that some of the standard workflow tasks were already prefilled with translation memory matches. A check revealed that a human mistake in the configuration of the translation memories had led to this issue in five of the ten Czech standard workflow tasks, in five of the Romanian standard workflow tasks, and one of the Polish standard workflow tasks. While it was still possible to reconfigure the Czech tasks, the affected Romanian and Polish tasks had already been completed by the responsible translators, who didn't notice the issue, and those tasks were therefore excluded from the data analysis, see Appendix A.1 for details.

As a result, the number of words varies between different languages and task types, and is not the same in terms of size and content. However, the average seconds per word values allow for comparison despite these variables.

Cochrane obtained the average seconds per words values by PLS and by language by using the Average secs / words values provided in the MateCat editing log, not by dividing Time-to-edit by Words. Calculating the values by dividing Time-to-edit totals by word totals, would lead to slightly different values in some cases, most likely due to rounded string-level values being used.

In addition, the participating translators as well as the selected content constitute a variable in the experiment. It is very unlikely that two translators will make exactly the same edits or produce the exact same translations, and different translators will likely work to different standards. Since Cochrane PLSs are specialized, cover a vast range of topics, and are hardly ever produced by the same authors, their terminology and linguistic complexity can vary a lot, and may be more or less challenging to translate. The variations in terms of PEE and average editing time per word across the different tasks are reflective of the translator and content variables.

3.3 Conclusions

The results from the post-editing experiment largely confirmed, but did not improve on the results from the small pilot conducted in June 2017: post-editing of final Y3 HimL MT outperformed Cochrane's standard translation workflow for three of the four HimL languages in terms of average time needed for editing, and subjective translator feedback suggests that translators not only saw the benefits of post-editing HimL MT compared to Cochrane's standard workflow, but mostly also accepted and preferred that work style. The less positive results for Polish post-editing of HimL MT are in line with previous evaluation tasks whereby Y3base HimL neural MT was not performing as well for Polish as it did for the other HimL languages.

Cochrane will be able to apply these findings to reconsider their standard translation workflow for selected languages, as it has become clear that post-editing of state of the art, neural, domain-adapted machine translation can substantially reduce the time

Annot.	Which language did you translate into?	Did you prefer to translate from scratch, or post-edit HimL machine translations?	Which approach do you think was quicker for you?	How would you rate the HimL machine translations for the purpose of post-editing?	Do you have any further comments?
1	Czech	Post-edit HimL machine translations	Post-editing HimL machine translations	Acceptable/ helpful	I would say the algorithms have much improved since the last post-editing round, definitely helpful when translating, very efficient and time-saving.
2	Czech	Translate from scratch	Not sure	Acceptable/ helpful	
3	German	Post-edit HimL machine translations	Post-editing HimL machine translations	Acceptable/ helpful	The post editing task was considerably easier and quicker.
4	German	Translate from scratch	Neither	Acceptable/ helpful	
5	German	Post-edit HimL machine translations	Post-editing HimL machine translations	Acceptable/ helpful	I think the machine translations worked very well!!!
6	German	Post-edit HimL machine translations	Post-editing HimL machine translations	Acceptable/ helpful	The HimL machine translation was far better than the automatic translation in Smartling and it would be great to use it for translations in the future.
7	Polish	Post-edit HimL machine translations	Post-editing HimL machine translations	Acceptable/ helpful	The difficulty depends on the PLS quality. Sometimes you just don't know what the authors wanted to say, sometimes sentences are long and I have to think twice how to express the idea in Polish. I would say that the PLS design can make problem to the machine translator.
8	Polish	Not sure	Post-editing HimL machine translations	Acceptable/ helpful	
9	Polish	Translate from scratch	Translating from scratch	Not acceptable/ not helpful	Polish is specific language and therefore machine translations are not very good.
10	Romanian	Post-edit HimL machine translations	Post-editing HimL machine translations	Acceptable/ helpful	

Table 10: Translator feedback survey results

needed to produce health translations, that translators see the benefits of the approach, and are willing to adapt their way of working.

4 Cochrane User Survey

The aim of Cochrane's user acceptance testing is to determine whether the HimL machine translations are of a high enough standard, despite perhaps containing errors, to be useful to Cochrane users reading them on the cochrane.org website.

Cochrane conducted anonymous user surveys via its website:

(a) from March to May 2017 evaluating the HimL Y2 MT engines,

(b) from mid-June to mid-August 2017 evaluating Y3base (referred to here, and in D5.5 as HimL Y2 neural) MT engines, and

(c) from mid-September to mid-December 2017 evaluating HimL Y3 MT engines.

Results from a) as well as preliminary results from b) were reported in D5.5. Final results from b) and c) are reported below.

4.1 Survey design and display adaptations for third user survey

The overall survey design and display followed that described in D5.5 (5.1 Survey design and display). For the final user survey evaluating HimL Y3 MT engines, adaptations were made as follows.

Cochrane increased the number of Plain Language Summaries (PLS) that were included in the experiment, aiming to attract a broader readership and potential survey participants (see more details in 4.2. below), as the earlier survey results had suggested a potential bias in responses due to an internal and high-level English-speaking readership. For Czech and Romanian, 1000 PLS were selected, machine translated and published as described in D4.2/5 (4.1.2 Publication and Display of Translations on Cochrane Website). For German and Polish, 100 PLS were selected, machine translated and published.

The survey consisted of the following questions and was translated into the four HimL languages:

1. "The translation below was generated using machine translation software. How easy is it to understand?" Users were asked to rate the translation from 1 star for very hard, 2 stars for hard, 3 stars for neutral, 4 stars for easy, to 5 stars for very easy. This question remained unchanged from the previous survey.

2. "Is this translation more useful for you than only seeing the original English text?" Users were asked to select yes or no, or could skip the question. This question remained unchanged from the previous survey.

3. "If you prefer to read the English version, please explain why." Cochrane decided to add this third, open-ended question, in a bid to clarify the answers given by respondents to question 2 that had proved inconclusive in the previous survey iteration.

As per the previous survey iterations, to collect comparison data, Cochrane added a similar pop-up survey to 100 randomly selected PLS that had previously been translated into German and Polish by Cochrane's volunteer teams.

4.2 Survey promotion

As before, Cochrane promoted the survey via its social media accounts and newsletters (a weekly communications update to about 100 Cochrane communicators around the world, monthly Cochrane Community newsletter, regional newsletters in local languages) and its Community website.

However, overall, Cochrane put less emphasis on promoting the survey within its network than it had for the first and second survey iterations, in an attempt to obtain responses from external respondents, with potentially less health-research and English-language proficiency as people from within Cochrane's community would typically have. The machine translations were displayed on cochrane.org in the same way as Cochrane's volunteer translations, and were indexed to appear in search engine results. The chosen PLSs were selected based on the top accessed topics by users from Czech Republic, Germany, Poland and Romania in the previous 12 months on the Cochrane website, as well as top search terms that had led such users from Google search to the Cochrane website in the past 90 days – information which can be obtained via Google Analytics – as PLSs linked to those popular topics and search terms were likely to attract genuine users via search engines.

4.3 Y2 neural MT user evaluation

The full analysis of the second survey evaluating Y2 neural HimL engines largely confirmed the preliminary results reported in D.5.5, suggesting that the quality of machine translations had improved compared to the first survey evaluating Y2 HimL engines for Czech, German and Polish, but not for Romanian, as detailed in Table 11.

Compared to the first survey, the median and mode ratings for German and Czech improved from 2 to 3 stars. The German mean rating increased from 2.1 to 3.0 stars, and the Czech mean rating from 2.2 to 2.7 stars. For Polish, the mean rating improved from 1.9 to 2.4 stars, the median rating remained at 2 stars, and the mode rating increased from 1 to 3 stars.

The Romanian results were slightly worse compared to the first survey, with the mean rating down from 1.9 to 1.6 stars, and mean and mode ratings remaining at 1 star, which was most likely related to an issue produced by the Romanian neural MT engine that had led to random sentences being added that did not match the source at all, as previously explained.

Less than 50% of German and Czech respondents said "yes" to machine translations being more useful than only seeing the English text, and for Polish and Romanian, a clear majority answered "no".

Language	Number of responses	Mean	Median	Mode	Yes, more useful	No, not more useful	Question 2 skipped
Czech	49	2.7	3	3	18	23	8
German	146	3.0	3	3	68	75	3
Polish	65	2.4	2	3	1	61	3
Romanian	219	1.6	1	1	38	127	54
All	479	2.3	2	1	125	286	68

Table 11: Y2 neural machine translation survey results: a) Mean, median and mode star ratings by language. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy. b) Responses by language whether the machine translation was more useful than only seeing the English text.

As before, the translations produced by Cochrane's Polish and German volunteers yielded very similar scores, with the average rating falling between easy and very easy to understand, and the median and the mode ratings being 5 stars for both languages, as shown in Table 12.

For the second survey question, the majority of Polish respondents did not find the translation more useful than only seeing the original English, while the German responses were more evenly split between "yes" and "no". These results do not seem to reflect the high comprehension scores awarded to the volunteer translations in question one. It has been suggested that this may have been due to a majority of the respondents speaking a very high level of English, and therefore didn't feel they needed a translation in their language, especially if it is not perfect. For the third user survey evaluating Y3 Himl MT, Cochrane attempted to clarify that assumption via an additional question as indicated above.

Language	Number of responses	Mean	Median	Mode	Yes, more useful	No, not more useful	Question 2 skipped
German	48	4.5	5	5	23	22	3
Polish	24	4.8	5	5	4	18	2
All	72	4.6	5	5	27	40	5

Table 12: Volunteer translation survey results, second survey: a) Mean, median and mode star ratings by language. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy. b) Responses by language whether the volunteer translation was more useful than only seeing the English text.

4.4 Y3 MT user evaluation

The third and final user survey evaluating Y3 HimL MT engines ran on the Cochrane website from 22 September until 18 December 2017.

4.4.1 Y3 machine translation survey results

A total of 177 responses were received for the machine translation survey across the four languages. That number was lower than in the previous surveys, which was probably due to the reduced promotion within Cochrane's own community, and the newly published machine translations taking some time to achieve high search engine rankings so that they can be found by users.

A summary of the survey results is presented in Table 13. Compared to the previous survey evaluating Y2 neural MT, the mean star rating increased from 2.7 to 3.6 stars for Czech, from 3.0 to 3.4 stars for German, and from 1.6 to 3.0 stars for Romanian. At mean ratings of or above 3 stars, the machine translations in those three languages were judged between neutral and easy to read on average, with highest ratings obtained for Czech. For Polish, however, the mean rating fell from 2.4 to 2.0 stars, and Polish machine translations received the lowest scores from respondents overall, rating them largely as hard to understand.

As before, looking at ratings of individual PLS there was variation between different PLS, different language translations of the same PLS, and there were wide ranges of ratings for the same PLS. For example, the Romanian machine translation of PLS CD011145 received ratings of 2, 4 and 5 stars, the Polish machine translation of PLS CD004467 was rated 5 times, with results ranging from 1 to 4 stars. The machine translation of PLS CD000526 was assigned ratings of 3 and 2 in Polish, but 3 and 5 in German. Generally, wide ranging ratings ranging for the same PLS were not uncommon.

Out of all 177 respondents, 40% said the machine translation was more useful than only seeing the English text, while just under half - 48% - responded that having the machine translation was not more useful than only seeing the English text. Looking at answers for specific languages, there was a sharp contrast in perceptions of usefulness of the German and Polish machine translation. In the case of German, 75% of respondents found the machine translation was more useful than only seeing the English text, while for Polish, only 6% said it was, reflecting again the poorly rated quality of the Polish machine translations. For Czech and Romanian, there were about the same number of "yes" and "no" answers to that question.

63 people provided an explanation as to why they had answered "no" to question 2 (Czech = 4, German = 5, Polish = 31, Romanian = 23). The most commonly given explanation was that the translation was of low quality, difficult to understand, unnatural, not fluid or contained incomprehensible terms or sentences. This feedback was particularly common for Polish, which echoed the low star ratings given. Other respondents said that they found the English original easier to understand, and more fluid, so this confirmed the earlier assumption that at least some of the respondents possessed good English skills.

The small sample of responses to the machine translation survey, in particular for Czech and German, could have biased the reported results.

Language	Number of responses	Mean	Median	Mode	Yes, more useful	No, not more useful	Question 2 skipped
Czech	16	3.6	4	4	8	6	2
German	33	3.4	3	3	25	6	2
Polish	46	2.0	2	2	3	36	7
Romanian	82	3.0	3	5	35	37	10
All	177	2.9	3	2	71	85	21

Table 13: Y3 machine translation survey results: a) Mean, median and mode star ratings by language. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy. b) Responses by language whether the machine translation was more useful than only seeing the English text.

4.4.2 Volunteer survey results

A total of 237 responses were received for the volunteer translation survey across Polish and German. This number was higher than in previous surveys probably due to the increased number of PLSs and the analytics-based selection of PLSs.

The third survey confirmed the good user acceptance of Cochrane's volunteer translations published in Polish and German recorded in the previous surveys, as shown in Table 14. The translations were mostly rated as easy and very easy to understand, with a mean score of 4.5 stars for German and 4.6 stars for Polish. 82% of German and 70% of Polish respondents found reading the PLS in Polish or German preferable to only seeing the English text, which largely echoes the high scores awarded to the volunteer translations for both languages. This result had improved compared to the previous survey, where high rates of "no" replies to question 2 seemed to contradict the high star ratings.

9 people provided an explanation as to why they had answered "no" to question 2 (German = 6, Polish = 3). The most commonly given explanations were that respondents were used to reading in English, found it more fluid or easier to read, or, in one case respectively, that they preferred to read the original, or the full text article instead of just the PLS.

4.5 Conclusions

Although Cochrane's volunteer translations continued to outperform the HimL machine translations in terms of user acceptance, the Y3 HimL engines obtained better user ratings than previously tested systems for Czech, German and Romanian. With up to 75% of German respondents and about 50% of Czech and Romanian respondents preferring the HimL machine translations over English text only, the results suggest that a substantial number of users would benefit from the provision of machine translated

Language	Number of responses	Mean	Median	Mode	Yes, more useful	No, not more useful	Question 2 skipped
German	171	4.5	5	5	140	10	21
Polish	66	4.6	5	5	46	8	12
All	237	4.5	5	5	186	18	33

Table 14: Volunteer translation survey results, third survey: a) Mean, median and mode star ratings by language. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy. b) Responses by language whether the volunteer translation is more useful than only seeing the English text.

Cochrane content, most likely those who are less comfortable reading in English. However, the results also suggest that the quality of the obtained machine translations can vary from one text to another. This could be due to the nature of the tested Cochrane texts, which cover a variety of topics, some more complex or specialised than others, and which also vary regarding complexity of used language, e.g. in terms of sentence length, grammatical structures and technical terminology. Standardisation and simplification of the source content could improve the usefulness of MT for Cochrane content in those languages further.

The lower user acceptance ratings for Polish demonstrate that the language barrier is more challenging to overcome via MT for some languages than for others, which Cochrane will further need to take into account when considering the use of MT to make its content available in different languages.

5 NHS24 User Survey

The user survey was developed on Survey Monkey and consisted of 2 parts. The first part asked how people access and use health information and NHS services. The second part included links to a test website with health information from www.nhsinform.scot translated using the HimL Y3 translation engine. Respondents were asked to access the site and give feedback on the usefulness of online health information, errors noted and quality of the translations. The surveys were professionally translated into Romanian and Polish.

The survey was launched in August 2017 and was disseminated to a wide range of third sector and public sector organisations who support the Polish and Romanian communities. This included Glasgow Health and Social Care Partnership, Aberdeen Polish Association, NHS Scotland Territorial Health Boards, Polish Schools and the Polish Family Support Centre.

The survey was closed on 30th November 2017. At that time, the Polish User Survey had been completed 46 times. The Romanian User Survey had been completed only twice. The results of the Polish and Romanian user Survey are shown in Table 15 and Table 16.

5.1 How people access health information

46% of Polish respondents had previously used the internet to access health information. When asked where they had accessed this information, respondents were given the option to select more than one source. 100% of respondents had used NHS websites; 36% had used Google. It is unclear whether the NHS websites were accessed via a search engine or directly.

Of the 2 responses to the Romanian survey, 1 (50%) had used the internet to access health information and the other had not. Every other question was skipped and therefore no further data was collected.

5.2 How people use NHS services in Scotland

Only 10% (3) of respondents reported that they had found it very easy to access health information in their own language while in Scotland. 24% of respondents reported that they had found it very difficult. 17% (5) of people reported using friends/family to translate for them and those who used a translator or interpreter. 52% (15) reported using Google Translate to access health information. Only 3% (1) of the respondents reported using a telephone interpretation service.

5.3 Translations

Of the 19 who answered the question "Would you recommend the website to people in your community?" 63% of respondents reported that they would recommend the HimL translations to friends and family. However 63% of respondents reported that the translations were not accurate. Looking at the responses to these two questions together shows that there is a need for translation and respondees would use what is available although they are aware that the quality of translations are not at a high standard.

The following feedback was given regarding the translations:

“For the person who does not know English the website will cause huge confusion. For someone with intermediate English that would be a jigsaw and the person should manage to figure out the meaning I recommend you to obtain a proper translation from interpreter to compare the outcomes.”

“The language is awkward and difficult to follow”

“Overall it has a very poor quality, very often sentences does not make sense. The words used are not correct, the grammar is very poor or totally wrong. I have a feeling that sometimes people would have difficulties in understanding the meaning of such translation. General theme - simple sentences generally are translated in acceptable way and could be understood, but the longer a sentence is (compound sentences) the translation is getting worse and is more difficult to understand. Sometimes the translation is done word by word and lacks the overall sense, meanings, each word is translated separately and does not match (express) the context of a sentence.”

“It's no yes or no answer. Some translations are so bad that they're confusing and misleading. Nevertheless, for people who don't speak English at all this could be some sort of help. There's a high chance they wouldn't understand Polish translation, but they will feel a bit more calm and they'll appreciate the gesture. Personally, I'd warn people I recommended the service to that there are lots of grammar and style mistakes.”

Question	Answer	Number of responses	% Total
Have you used the internet to access Scottish Health Information before?	Yes	21	45.5
	No	25	45.5
How did you access this information? Tick all that apply	NHS websites	11	100.0
	Search Engines	4	36.4
	Other Websites	1	9.1
	Twitter	0	0.0
	Facebook	2	18.2
How difficult has it been to access health information in your own language while in Scotland?	Very Easy	3	10.3
	Easy	4	13.8
	Moderate	10	34.5
	Difficult	5	17.2
	Very Difficult	7	24.1
Have you used any of the following translation aids when using online or face to face NHS services?	Translator or Interpretor	8	27.6
	Google Translate	15	51.7
	Telephone interpretor	1	3.4
	Friends or family	5	17.2
	Other	10	34.4
How good is the quality of the translations overall?	Very Good	1	5.3
	Good	9	47.3
	Unsure	4	21.1
	Poor	4	21.1
	Very Poor	1	5.3
Would you recommend the website to people in your community?	Yes	12	63.1
	No	7	36.8

Table 15: Polish Survey Results

Question	Answer	Number of responses	% Total
Have you used the internet to access Scottish Health Information before?	Yes	1	50.0
	No	1	50.0

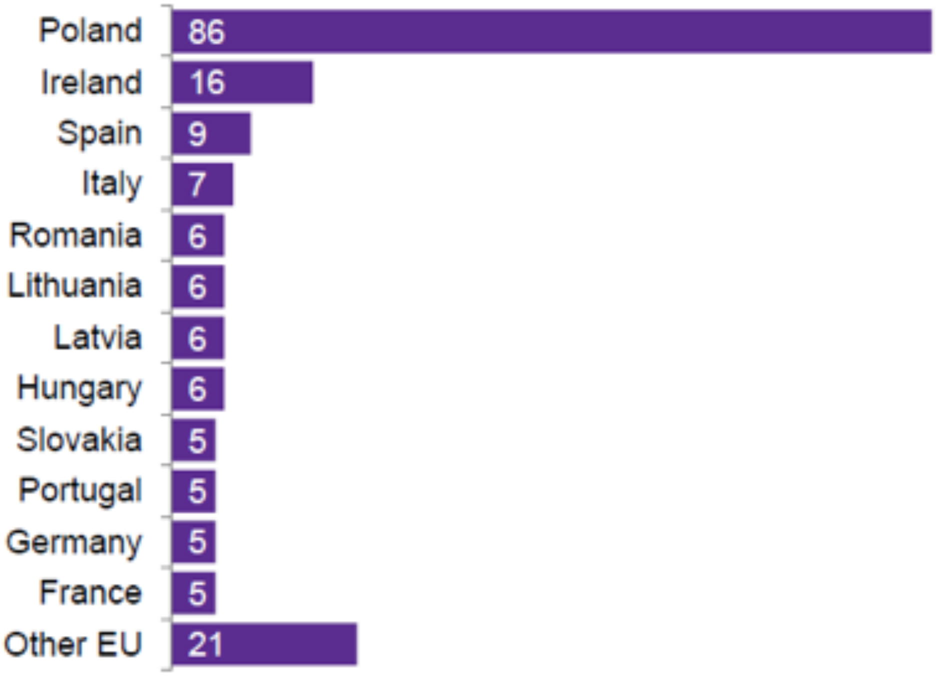
Table 16: Romanian Survey Results

6 NHS24 Needs Impact Study

In this section we describe two needs impact studies performed for the NHS24. In-depth interviews were conducted with Polish speakers and Romanian speakers who were potential NHSinform translated website users in order to determine what information needs they have, and if and how machine translation should be incorporated into NHS24 services.

6.1 Overview

There is a sizeable population in Scotland with cultural and linguistic barriers to providing health information. In Figure 2 we can see statistics provided by the Scottish Parliament Information Centre report 16/86, which gives the numbers in 000's of EU nationals by country of origin and so an indication of the size of the Scottish population who may be interested in MT into the HimL languages. The number of Czech nationals in Scotland is reported as 2,245 in the 2011 Census. The SPICe report indicates that this has not risen to more than 5,000 in 2015.



Source: SPICe analysis of APS Jan-Dec 2015, ONS

Figure 2: Numbers by country of origin

Translations of NHSinform content help users to better understand health information and this will assist in achieving equality of access to health services. Better equality of access leads to better health for the Scottish population as a whole. Manual translation of all health information content into all languages present in Scotland is impractical; the number of languages present in the resident population and the change in understanding of conditions and their treatment and changes to the provision of services make translation by human translators too expensive.

NHS 24 has adopted two policies in recent times;

1. translate most popular information on NHSinform into the languages most present in Scotland
2. provide NHSinform content in Plain English, suitable for a reading age of 9 years, and translations only on request.

The difficulty and cost of keeping up with changes to information content and knowing which content to translate into which languages resulted in the adoption of the plain English and translation on request policy. NHS 24's interest in the HimL project was to have a more readily available and cost-effective translation system for health information.

The objectives of the Needs impact study were to;

- Identify the user need for NHS Scotland health information translated into target languages

- Establish how accurate the MT translations were for the target populations
- Determine the likelihood of using a website for health information
- Identify how to raise awareness of the site

Respondents were recruited on the basis that they;

- Could not speak or read English at anything other than a basic level
- Were parents of children under 11

6.2 Method

In-depth interviews were conducted with 22 Polish speakers and 17 Romanian speakers who were potential NHSinform translated website users. In addition, 6 in-depth interviews were conducted with community group workers. Interviews lasted 45-60 minutes. The fieldwork was conducted between 24th May 2017 and 15th June 2017. An incentive of £40 gift voucher was given as a thank-you for participation. Interviews were conducted in Glasgow, Edinburgh and Aberdeen.

Community support workers and users were interviewed, using an interpreter where necessary, to get their views of the translations and their approaches to accessing health information. Oral Health, fever in Children and Chickenpox were taken as providing popular and relevant source content from NHSinform and translated using the Y2 phrase-based MT system and uploaded to a test site, which displayed both the English and translation (on separate pages). Interviews were conducted using two separate topic guides, see Appendix A.1, with reference to the test site.

About 6 months later, the Y3 system, which uses AI techniques to translate English, was also reviewed, using the same support worker subjects. The same NHSinform content as before was translated and uploaded to the test site. The topic guide for these interviews is given in Appendix A.3.

The Y3 system was also reviewed by 6 community group worker subjects, 5 of whom had participated in the earlier interviews and 1 who had not previously seen the translated site. The same NHSinform content as before was translated and uploaded to the test site.

6.3 Results

6.3.1 User Needs

All literate respondents liked the concept of health information online, translated into their language. Regular users of online services reported that they used Google translate but were aware of its limitations. Others used Polish or Romanian sites as a source of health information.

To make a translated website worthwhile according to these communities, it needs to;

- Be translated accurately so there is no ambiguity
- Have the NHS logo on it to add credibility
- Be promoted as a 'go to' health site
- Be promoted amongst the younger community, as they are more likely to have online access and be proxy users for family
- Have printed versions for those who cannot access the internet
- Be promoted and trusted by community representatives so they can discuss anything the audience don't understand

There is a need for health information in Romanian, however a website alone is not inclusive enough. There is a need for audio versions for those who are illiterate and potentially more in need of the information than other community members.

6.3.2 Translations

The translations were deemed to be of poor quality and it was obvious to respondents from the outset that they had been machine translated. The translations were reported as being no better than Google translate, which most were familiar with. Respondents could work out what the site was telling them but;

- They had to put in a lot of effort to read sentences to make sense of them
- Some asked the interpreter present to confirm a word was incorrect
- Some misunderstood what was being said
- Incorrect words were used which in some cases was the opposite of what was intended

Respondents identified some common errors in the Y2 translations;

- Incorrect words being used in translations
- Words missing from translations
- Meaning of sentences not being clear from the translation, sometimes with instructions being the opposite of the instructions in English

6 support workers were re-interviewed with the Y3 system translations and reported that;

- The Y3 translations of words and phrases were an improvement over Y2 but there were still some single-word mistranslations
- The Y3 grammar was better
- The Y3 translations read more naturally and full sentences made sense
- The general understanding of each page of content was better
- There were no instances of opposite advice being given when compared to the English

6.3.3 Likelihood of use

There was concern over using the site with the errors identified. Respondents said they would use such a site if the translations were more accurate. The test site did not provide a big enough benefit to be used.

There was more general concern over knowing which online health information is trustworthy. Will use Google translate for words and phrases.

Those who could read a little English stated that they would use the English in conjunction with the translation;

- To ensure their understanding was correct by comparing the two versions
- To use Google translate for words they were unsure of
- To use the English version to explain their concerns to their GP

The improvements in translations present in the Y3 system meant that this was more likely to be recommended by support workers.

6.3.4 How to raise awareness

None of the interviewees were aware of NHSinform before the interviews. Some respondents would use the website now they were aware of it but would need to translate the content. Most liked the concept. Romanians rely on their community connections and will support and promote ideas to each other. Some Romanians have been disengaged from the wider society for all their life and so know very little about the Health service or what rights they have. Word of mouth is important to both communities.

6.4 Discussion

6.4.1 Romanian speakers in Scotland

It should be noted that the Romanian community in Scotland has 2 distinct components: "European" Romanians and members of the Roma community. The study focused on the European Romanians as this was one of the target languages for the project. There are a larger group of Romanian immigrants to Scotland who are well educated and are here to study and work and have a good level of English. There is a much smaller community who have a poor level of English, the majority of this group are illiterate.

Of the respondents who were interviewed some were completely illiterate and could not write their own names. One couple were highly dependent on their 12 year old daughter to translate for them and to use the internet for information. Another couple who could read and write Romanian would access Romanian websites for information or use Google translate.

6.4.2 Polish speakers in Scotland

All Polish respondents were literate with a mix of ability in speaking English and would make some attempts to read English. All were familiar with the internet and used a range of devices.

There are differences in expectations of the health service in Poland and Scotland. Users expect antibiotics on request in Poland but sometimes are only given paracetamol in Scotland. The role of nurses differs in Poland they are much more like care assistants than medically qualified professionals.

6.4.3 Use of NHS Scotland Services

All respondents were heavily dependent on GPs and used them as their first port of call for health services. Interpretation services are always available. Face to face support gives peace of mind, particularly in respect of children's health. Even if information has been sought online, many will often book an appointment to get reassurance from their GP. Google translate is used to refine searches and to translate prescriptions and letter from the NHS. One suggestion was to include content about the differences between the health services in Scotland and Poland. The NHS brand is recognised and respected but the accuracy of translations in this research was a major disincentive to use. There is the expectation that advice on an NHS website would be 100% accurate.

6.5 Conclusion

The conclusion of this study, backed up by other user testing, is that the translation errors make publication of unrevised machine translations unacceptable. Community workers report that although the later Y3 release showed clear improvements, significant errors still existed in the translations. These could be potentially life-threatening and damaging to the NHS brand and reputation if translations were not at least as good as by professional human translators. To do this with machine translation requires post-editing by a professional human translator before publication on their sites.

7 Impact: Web Traffic

7.1 Analysis of Cochrane web traffic related to HimL languages and countries

In D5.5 (8. Impact of translations on Cochrane web traffic), Cochrane had reported on the impact of translations on its web traffic in general, and provided some initial analysis of web access specific to HimL languages and countries to date.

Following that, Cochrane continued to monitor web traffic to evaluate the effect of further evaluation and dissemination activities: the publication of additional HimL machine translations (MT), running the user survey again for the Y3 systems, and dissemination efforts in the HimL languages. The results are presented below.

7.1.1 Context of the final web traffic analysis

In September 2017, around 1,000 Cochrane Plain Language Summaries (PLS) were machine translated into Czech and Romanian using the final Y3 HimL MT engines, and added to cochrane.org as part of the user survey evaluation, and on the premise that having more content available to users of these languages would result in increased access by genuine users.

Approximately 100 PLS were machine translated into Polish and German and added to their respective language versions of cochrane.org as part of the user survey, but it was expected that the effect on access would be less notable and hard to measure, given the existing human translations in those languages that have been published and disseminated for several years. The HimL machine translations only accounted for about 7% of all German language content, and 11% of all Polish language content respectively.

In general, search engines take some time to index new web content, and there are numerous factors affecting whether web content is ranked high or low in search engine results, which are hard to predict, so it wasn't clear what effect the additional translations would have overall, in the relatively brief time frame of the analysis.

Nonetheless, the analysis below will compare the first six months of 2017, from 1 January to 30 June, with 1 July to 31 December, to assess the effect of adding the extra machine translated PLS, in addition to the dissemination efforts that were taking place.

7.1.2 Analysis of web traffic related to HimL languages and countries

As before, Cochrane assessed access by Internet users in Romania, Germany, Czech Republic and Poland, as well as visits by people using the Internet with browsers set to the Romanian, German, Czech and Polish languages. Some limitations of the relevant data from Google Analytics were described in D5.5. In addition, Cochrane looked at page views in different languages, i.e. how many users accessed pages in the HimL languages, as the country and language browser visits do not differentiate whether a user visits a German or English page, for example.

Visits by HimL countries: comparison of January - June 2017 with July - December 2017

As illustrated in Figure 3, there were more visits from people in Romania, Poland and the Czech Republic during the second half of 2017 than the first, while access from Germany was higher in the first half of 2017.

The most noteworthy statistic here is a 50.4% increase in visits by people accessing cochrane.org from Romania in the second half of the year, which could be attributed to having more information available in Romanian from September 2017 onwards. However, the same impact was not seen in the case of visits from the Czech Republic; there was no significant rise in visits, comparing 4,581 in January to June, with a slight increase to 4,611 from July to December.

The increase for Poland, and the decrease for Germany cannot be linked to the HimL translations specifically; as indicated above, for both languages new human translations had been published and disseminated for several years.

Access to cochrane.org by visitors using browsers set to HimL languages: comparison of January - June 2017 with July - December 2017

In a similar situation to that seen in Figure 3, Figure 4 shows there were more visits by browsers set to German in the first half of the year, while visits by browsers set to Polish increased in the second half of the year. Visits by users with Czech language browsers only slightly increased comparing the two halves of the year, with 3,112 visits in the first and 3,285 in the second. Romanian has seen the most encouraging increase, with visits more than doubling from 2,901 in the first half of the year, to 5,952 in the second half.

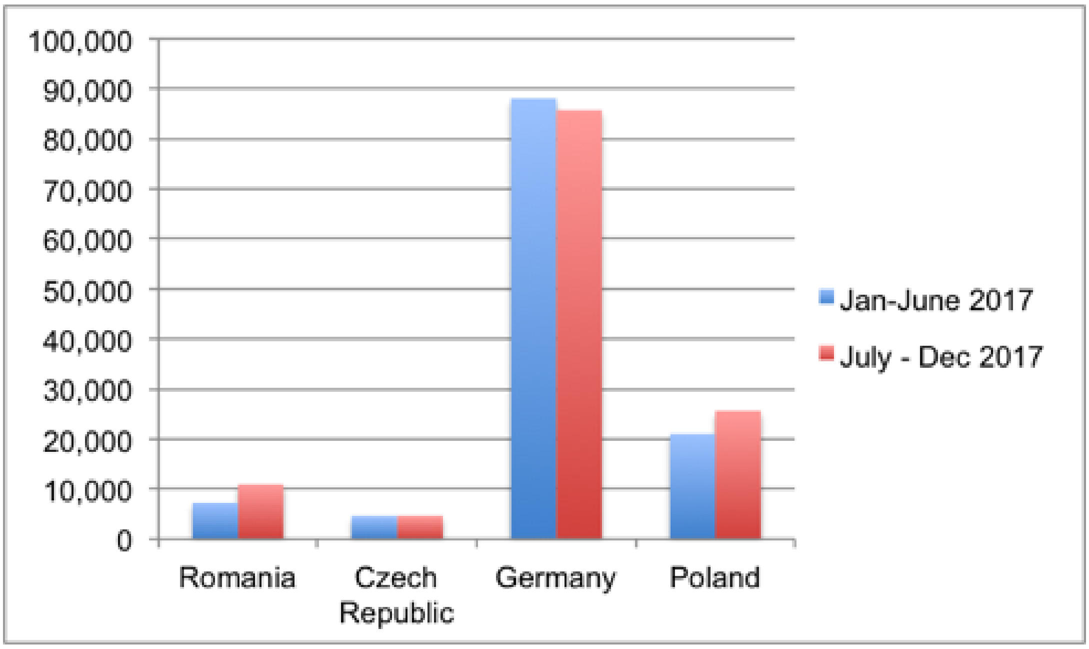


Figure 3: Number of visits to cochrane.org from HimL countries: January - June 2017 and July - December 2017.

Visits by Internet users browsing in Czech and Romanian broken down on a monthly basis in the line graph in Figure 5 show a sharp increase in access by web users navigating with a Romanian or Czech browser, following the introduction of more health content in the two languages in September 2017. In the case of Romanian, visits have risen steadily since making the extra content available, while Czech also saw a clear increase, although not as big, and experienced a small fall from November to December.

Figure 6 shows the contrast to visits by web users navigating with a German or Polish browser. Overall the numbers are higher than those for Romanian and Czech due to the history of the existing translations in German and Polish, and there is a less clear high towards the end of the year, following the addition of the Y3 HimL MT engine translations. In the case of Polish, visits have been consistently increasing nearly each month in 2017. There were 2,614 visits in January 2017, increasing to 5,294 in December 2017 – an increase of 102%. German has seen less consistency in terms of monthly visits, illustrated by several peaks throughout the year, with a high in November, correlating with the availability of the additional HimL machine translations. German has also seen more of a dramatic difference in numbers by month; for example, German web browsers accounted for 20,121 visits in November 2017, then falling by about 20% to 15,942 in December 2017. From Cochrane's experience, these fluctuations could at least partly be seasonal, as it is not unusual to have drops in access during Western summer and winter holiday periods. A common reason for a spike in traffic can be the publication and dissemination of a Review that sparks particular public interest, or a targeted media campaign, that would lead to a temporary increase in access.

Number of page views of cochrane.org content in HimL languages: comparison of January - June 2017 with July - December 2017

Figure 7 shows a strong increase in number of page views of content published on cochrane.org in Czech and Romanian, comparing the first and second halves of 2017. For Czech content, the number had more than doubled compared to the previous half year, and was almost five times higher for Romanian, which was facilitated by a lot more Czech and Romanian content being available on the site. So while the total number of visitors from Czech Republic and Romania, or visitors using Czech and Romanian browsers, hadn't increased dramatically from the first to the second half of the year, it would seem likely, based on the page views, that many more of these users accessed content in their language, as opposed to having to use English content.

The page views for German and Polish as shown in Figure 8 reflect the number of visits by country and browser language provided above: German hits decreased, Polish hits increased from the first to the second half of 2017.

Looking specifically at the number of page views obtained for HimL machine translations added to cochrane.org in September, there were 563 page views of Polish machine translations, accounting for just fewer than 2.7% of all Polish content page views during that period. Views of German machine translations, were even lower – 402 page views of machine translations, accounting for only 0.8% of all German page views. These numbers are likely a reflection of the strong performance of Cochrane's human translation and its dissemination activities built up in the past, rather than a reflection of low relevance of the HimL machine translations. So the machine translations didn't have a visible effect on substantially increasing web traffic in the context

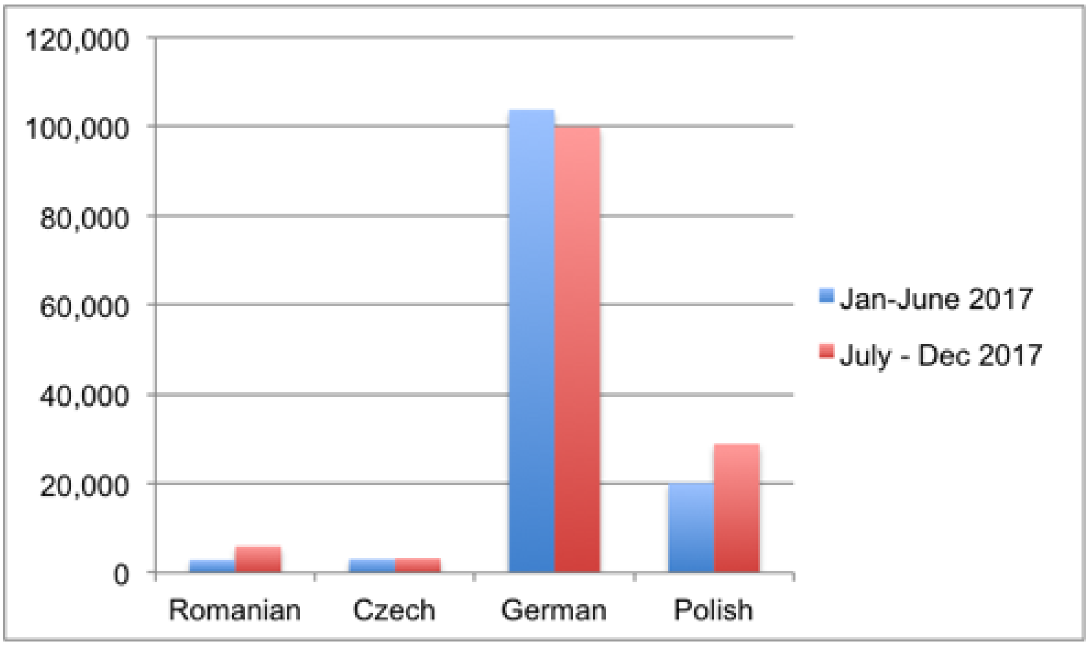


Figure 4: Number of visits to cochrane.org from users of browsers set to the HimL languages: January - June 2017 compared to July - December 2017.

of those two languages that already had relatively high access numbers, and human translations published and disseminated for several years.

7.1.3 Conclusion

The encouraging increase in access seen for Romanian users and content, and to some extent for Czech users and content, confirmed Cochrane's previous experience that making content available in different languages allows users of those languages to find the health information that they are looking for. The analysis further showed that this effect can be achieved with HimL machine translations, and not only human translations. Although the numbers were still relatively low for Romanian and Czech when compared to access for the other languages, this was to be expected, as Cochrane has experienced similar slow starts when adding other new languages to cochrane.org in the past – it takes time to build an audience and to rank well in search engines. It is likely that the access numbers for Romanian and Czech would have further increased over a longer period of time.

These results will allow Cochrane to consider the potential benefit of using domain adapted machine translation as part of their translation strategy, e.g. for languages where no human translators are available (yet).

7.2 NHS 24 Web Traffic

7.2.1 Background

NHSinform is Scotland's national health information service for the public. It contains information about illnesses and conditions, tests and treatments, healthy living advice, health rights, care and support services. It also has a self-help guide, webchat, text to speech and translation functions via BrowseAloud. There are also links to NHS, local authority and charity organisations providing advice and services. The site is written according to the AA standard of the Web Content Accessibility Guidelines (WCAG) 2.0 in Plain English aimed at a reading age of 11 years.

Users can request translation of any of the pages on NHSinform, either through the 'contact us' feature or in discussion with one of the health information advisors. Such requests usually come through an English-speaking family member. An interpreter is always available to anyone on the phone to a health information advisor or any other NHS 24 provided services. The service allows a three-way conversation to be held and takes less than a minute to establish.

Prior to the EU HimL project, there was an earlier NHS24 Health in my Language site which ran from 2011 to 2015. It provided the top 20 frequently asked questions in 5 most frequently requested languages; Polish, Mandarin, Gaelic, EasyRead and British Sign Language (as video clips). The NHS24 HimL site was established and refreshed once during its lifetime. It was not maintained partly because of a lack of funding for pro-active translations, partly because of low use (less than 1,000 visits per month) and partly because of the intention to use machine translation on a redesigned NHSinform website.

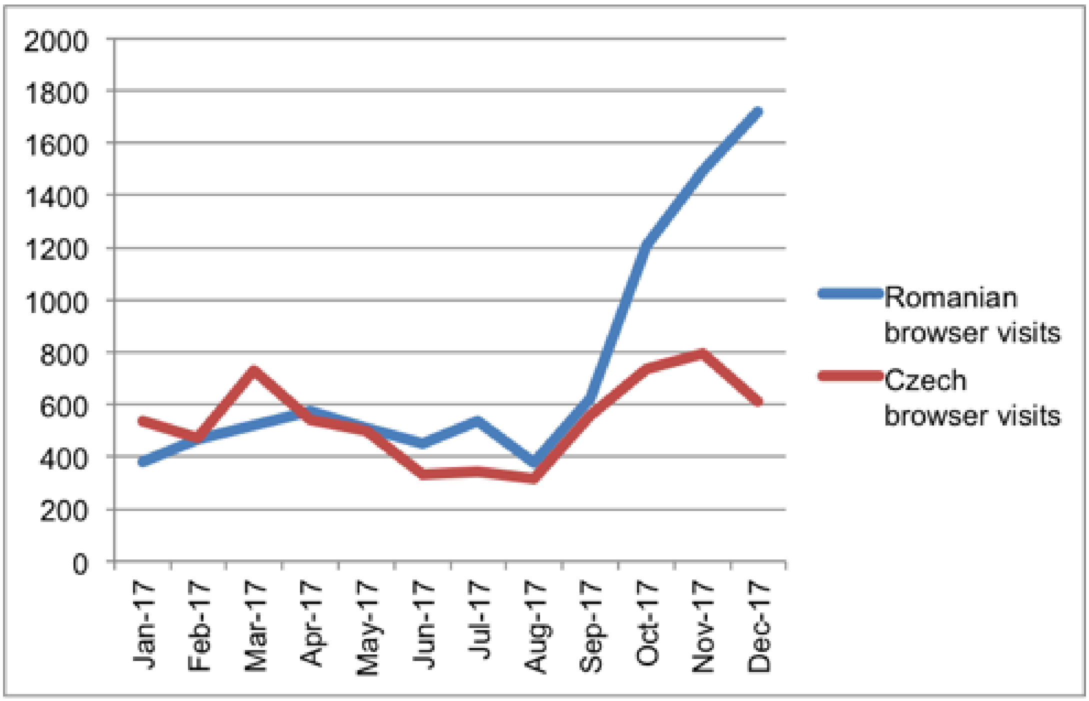


Figure 5: Monthly breakdown of number of visits to cochrane.org by users with web browsers set to Czech and Romanian from January to December 2017.

The EU HimL project translated and tested in 4 languages; Polish, Czech, German and Romanian, determined by the needs of the user partners and because they cover the three main European language groups (Slavic, Romance, Germanic).

At the start of the project it had been the intention to deploy Machine Translated content to NHSinform for both the Y2 and Y3 systems and observe the impact this had on visits to the site by country and by browser language setting. However, findings from the evaluation work stream indicated that the Y2 and Y3 MT were insufficiently accurate in a medical context to be published without manual translation. NHS 24 decided to continue testing the Y3 MT output but not to publish MT content and removed all MT content from NHSinform. This report focuses on the changes to visits to the NHSinform site by country and by browser language over the evaluation period.

7.2.2 Previous NHS24 HimL site

In this section we report the web site traffic for the previous NHS Health in My Language Website. These results show the low amount of traffic to the human translated content. Statistics were collected in May 2012 of visits to the NHSinform and NHS24 HimL sites and are shown in Table 17.

Language	NHS24 HimL	NHSinform
English	563	22593
German	2	15
Polish	13	24
Romanian	0	1
Czech	0	2
Total Visits	588	22780

Table 17: Visits by language to NHS24 HimL and NHSinform May 2012

Of visits to the NHSinform site 0.8% were from users with language settings other than English. Visits from users with the language set to one of the EUHimL languages were only 0.2% The NHS24 HimL site was prominently signposted on NHSinform, with the intention of directing non-English traffic to HimL.

In Table 18 we can see that of visits to HimL in May 2012, 13.4% were from countries other than the UK. Of the two visits from Poland, both had language set to Polish. Of visits to NHSinform in May 2012, 5.5% were from countries other than the UK. Germany was the EUHimL Country most represented, possibly due to the numbers of military personnel based there. Of the 38

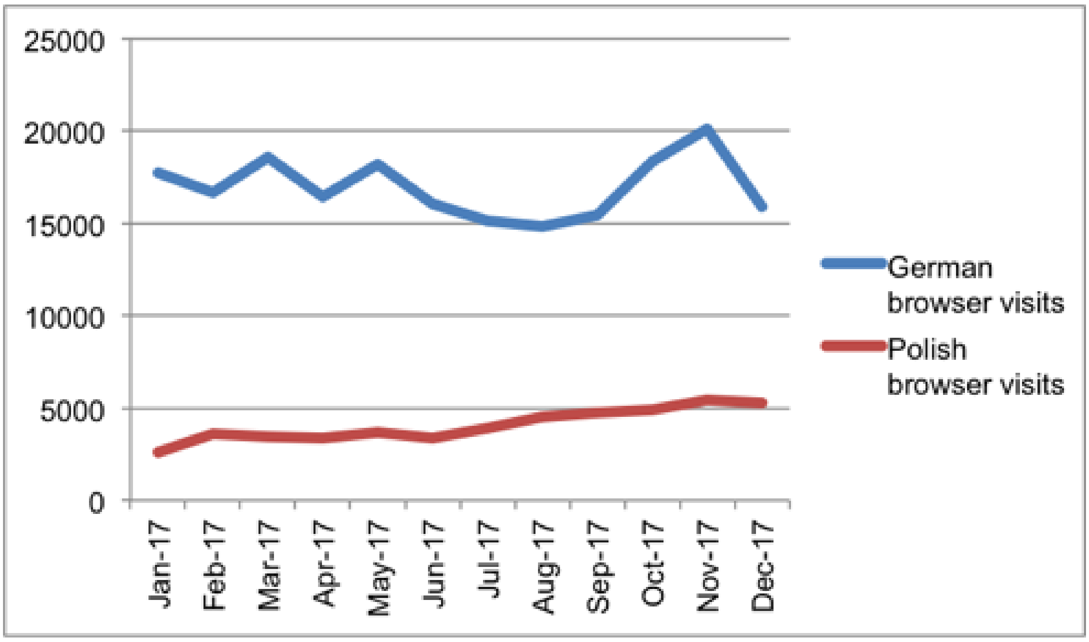


Figure 6: Monthly breakdown of number of visits to cochrane.org with web browsers set to German and Polish from January to December 2017.

Country	Visits to HimL	% of HimL visits by Country	Visits to NHSinform	% visits to NHSinform by Country	% HimL of NHSinform
UK	507	86.2 %	21531	94.5 %	2.3 %
Germany	0	0	38	0.2 %	0
Poland	2	0.3 %	2	0.01 %	100 %
Romania	0	0	2	0.01 %	0
Czech Republic	0	0	0	0	0
Other countries	77	13.1 %	1207	5.3 %	6.4 %
Total Visits	588	22780		2.6 %	

Table 18: Visits by country to previous NHS24 HimL and NHSinform May 2012

visitors from Germany, 30 had an English variant as their language setting, only 5 had German set. Of the visits from Poland, both had Polish as their language setting. Of the two visits where the language setting was Romanian, neither had a Country setting. In May 2012, the HimL site received only 2.6% of the number of visits of NHSinform. Redirect information was not recorded.

7.2.3 Approach

It had been the intention of the project to automatically deploy MT content onto the live NHSinform site but whilst redeveloping the NHSinform website during 2015 and 2016, but this was not possible. The design and implementation of the EUHimL deployment facility meant that content could not be translated on user request and therefore had to be pre-translated off-line with some means of navigation. Therefore an EUHiml test site was generated for user testing purposes. As the test site was only available to invited subjects its usage was not monitored. As a result of the evaluation of MT content from the test site, a decision was taken not to deploy MT to the NHSinform live site. In the rest of this section we therefore report website traffic for the original English NHSinform web pages, looking at their browser language and country of origin to gain insights about the language needs of our visitors.

The live NHSinform sites (both pre and post redesign) were monitored for access by country, determined by originating ip address, and by user browser language setting. All language variants (e.g en, en-us, en-uk) were recorded and summed. Countries and languages to be monitored were agreed with Cochrane. The statistics collected from Google Analytics were; Number of sessions, new users, pages / session, average time on site, and bounce rate. Analysis is reported on the number of sessions. No

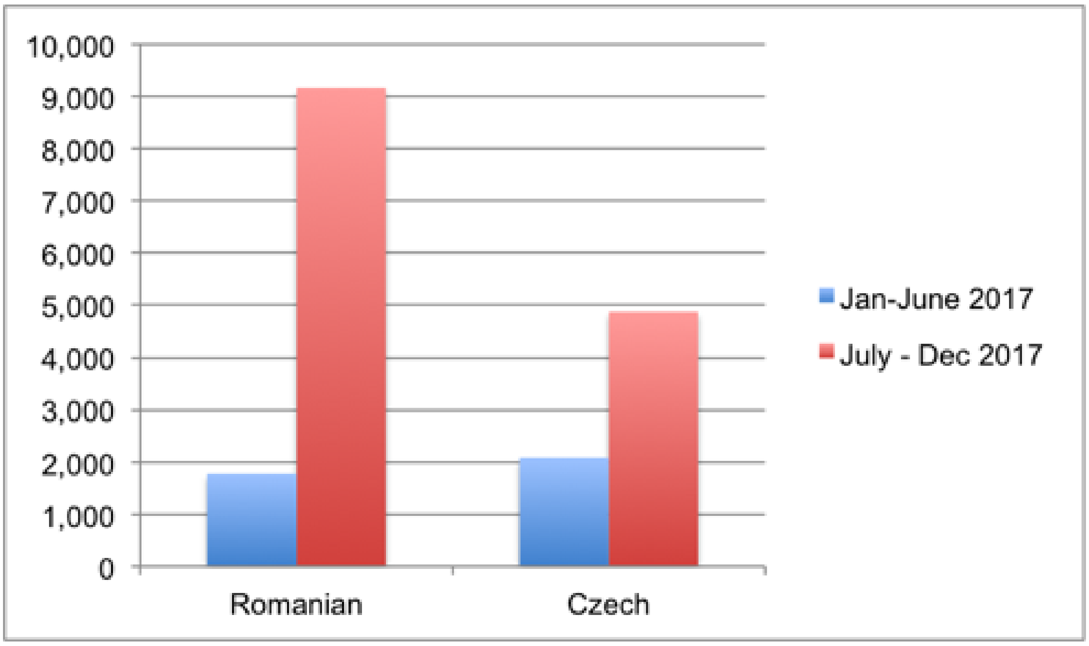


Figure 7: Number of page views on Czech and Romanian content of cochrane.org: January - June 2017 compared to July - December 2017.

attempt was made to identify bot access.

7.2.4 Analysis

In Figure 9 we can see the total number of visitor sessions to all NHSinform websites during period July 2016 and July 2017.

The NHSinform site showed an average growth during this period of around 900% (Figure 9). The NHSinform website shows an increase on introduction of the new site and a marked growth between April and July 2017.

As noted above, the NHSinform site was redesigned during this period and included access for mobile devices, previously provided by a separate site. 2016 was the old site and 2017 the new site.

As seen in Figure 10, the proportion of visits from HimL languages browser settings showed an increase of around 0.1% on the introduction of the new site. The apparent decline from April to July 2017 is due to the greater growth in use of NHSinform, rather than a decline in use by users with non-English language settings.

Visits by country

Country	Visits 2016	% of Total 2016	Visits 2017	% of Total 2017	% Growth 2017-2018
UK	31,200	88.09 %	224,996	70.94 %	721 %
Germany	77	0.22 %	709	0.22 %	921 %
Poland	32	0.09 %	172	0.05 %	538 %
Romania	21	0.06 %	181	0.06 %	862 %
Czech Rep.	9	0.03 %	90	0.03 %	1000 %
Top 9 Non-HimL Countries	2,810	7.93 %	70,373	22.19 %	2504 %
Total	35,418		317,168		895 %
HimL % of total	0.39		0.36		

Table 19: Visits by country to HimL and NHSinform May 2012

In this section we describe the number and percentage of visits broken down by the country of the visitors as determined by their IP address. Table 19 shows the visits to NHSinform from HimL countries, alongside the UK and top 9 non-HimL countries.

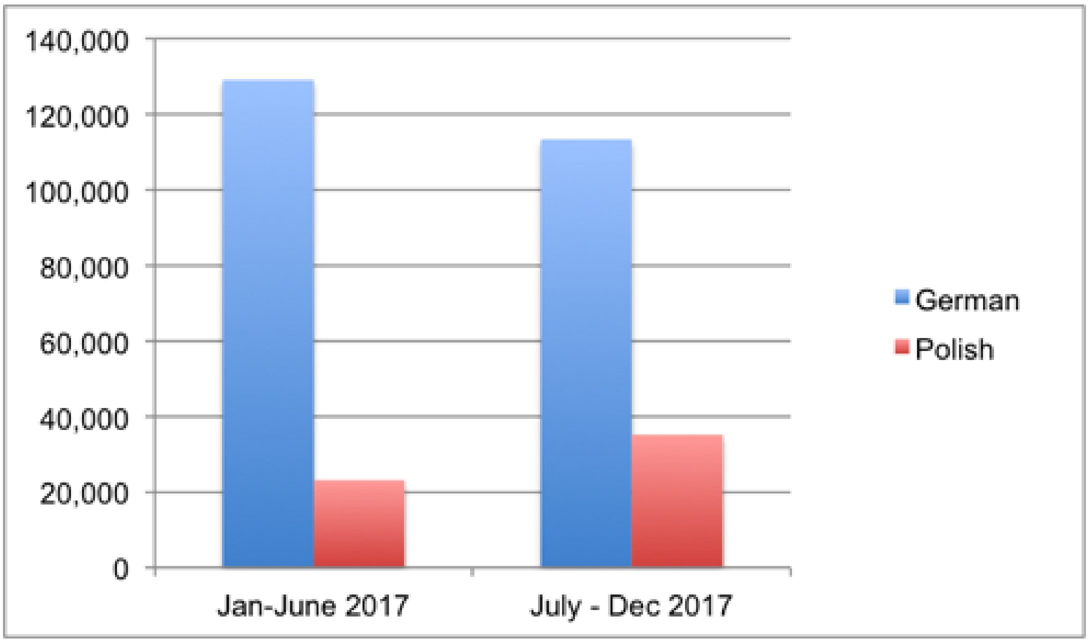


Figure 8: Number of page views on German and Polish content of cochrane.org: January - June 2017 compared to July - December 2017.

Growth is coming from all countries but more from outside Europe. This may partly be due to growth in computer use around the world and the inclusion of mobile access to the site.

In Figure 11 we see the visits to NHSinform by country show large growth from 2016 to 2017, however this must be viewed in context by looking at the growth in the proportion of visits by country, to take into account the overall growth in visits to NHSinform.

In Figure 12 we see that the proportion of visits from HimL countries remains relatively unchanged.

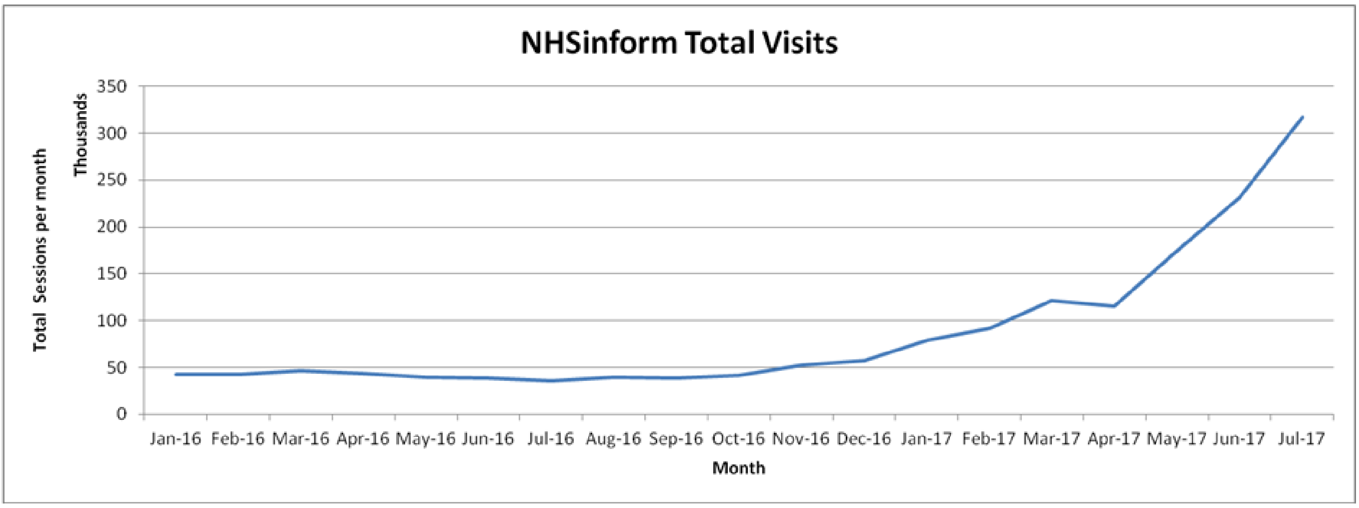


Figure 9: Visits to NHSinform by HimL language browser setting July 2016 to July 2017

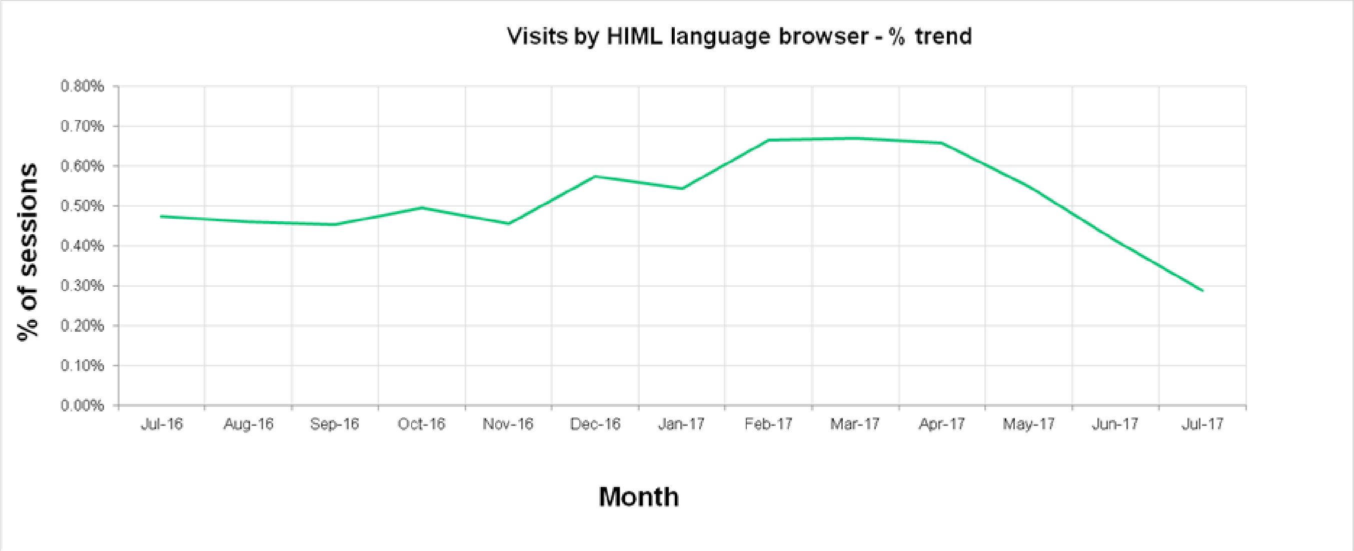


Figure 10: Proportion of Visits to NHSinform by HimL language browser setting July 2016 to July 2017

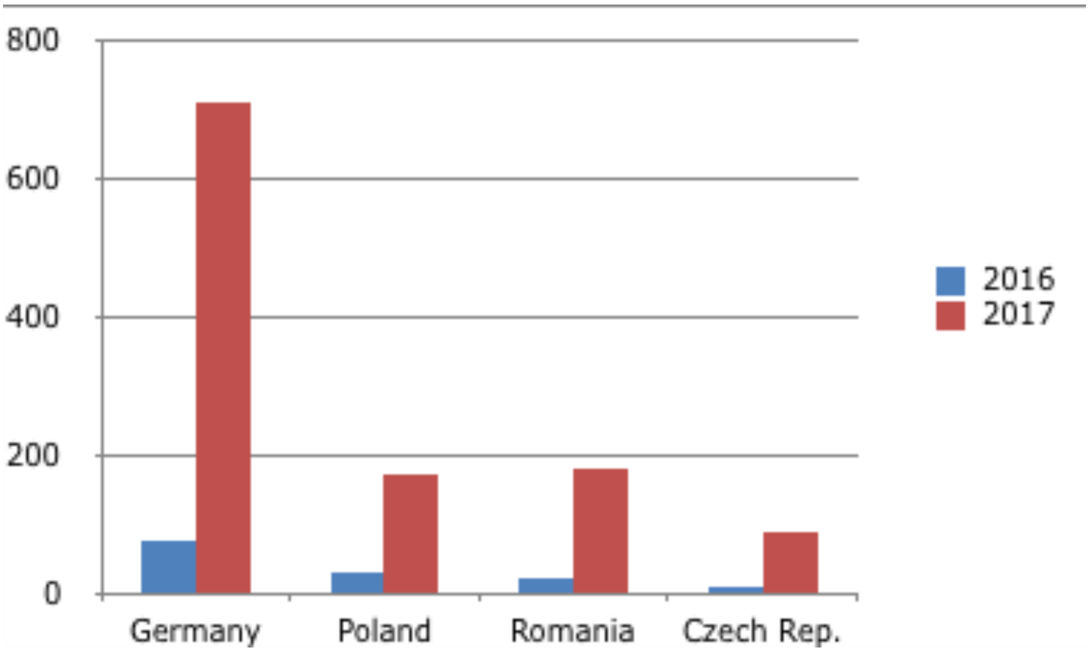


Figure 11: Visits to NHSinform from users of HimL countries January to July 2016 and 2017

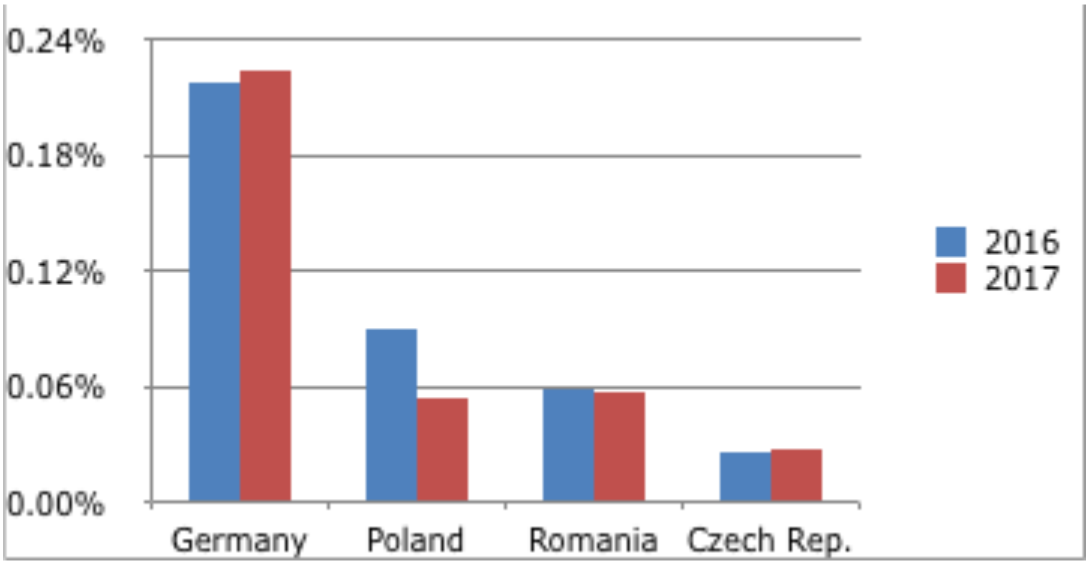


Figure 12: Visits by country as a proportion of total visits to NHSinform January to July 2016 and 2017

Visits by HimL language

Language/Year	Jan-Jul 2016	% of total 2016	Jan-Jul 2017	% of total 2017	% Growth 2016-2017
English	34660	98.78 %	311347	99.45 %	898 %
German	57	0.16 %	331	0.11 %	581 %
Polish	95	0.27 %	472	0.15 %	497 %
Romanian	10	0.03 %	70	0.02 %	700 %
Czech	6	0.02 %	41	0.01 %	683 %
Other Non-English (by difference)	259	0.74 %	822	0.26 %	317 %
Total	35087		313083		892 %
HimL % of Total	0.48 %		0.29 %		

Table 20: Visits to NHSinform by HimL Language January to July 2016 and 2017

In this section we describe the number and percentage of visits broken down by the language settings of the visitor's browsers. Table 20 shows the number and proportion of visits to NHSinform with different browser language settings January to July 2016 and 2017.

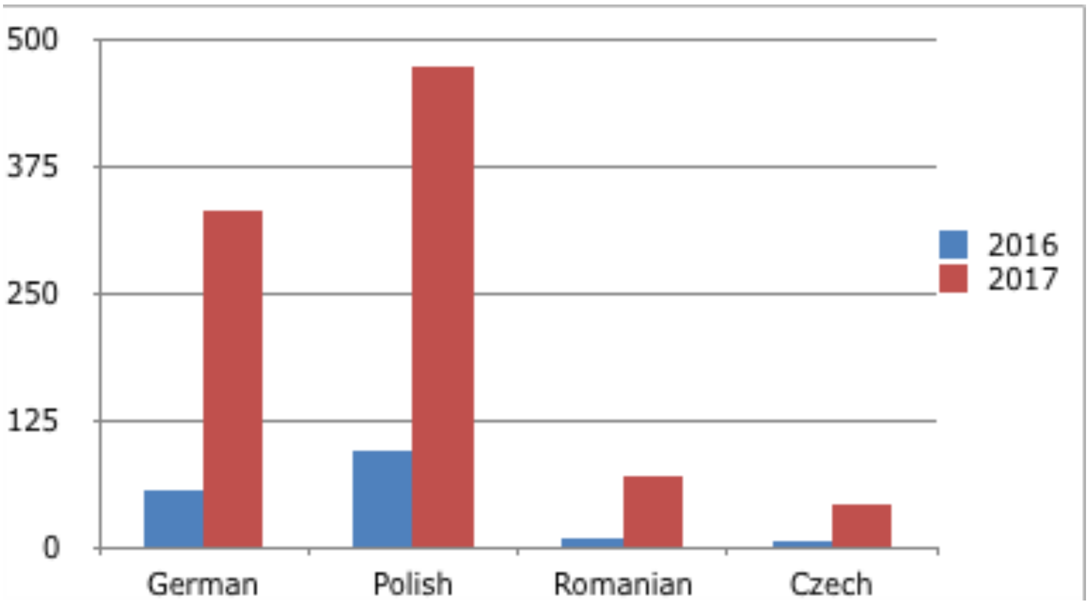


Figure 13: Number of visits to NHSinform by users of browsers set to HimL languages January to July 2016 compared to the same period 2017

In Figure 13 we can see that although the total number of visits by language increased dramatically from 2016 to 2017, the proportion of visits from HimL language users declined from 0.48% to 0.29% of total visits. The Needs Impact study reported in section 6 indicates that it is most likely due to a lack of awareness of NHSinform amongst target audiences.

The proportion of visits to NHSinform by non-English browser settings is only 0.48% compared with 66% non-English browser settings reported for Cochrane.org in 2016. This is likely to be a reflection of the difference in the target audiences for the two sites.

7.2.5 Outlook

Despite there being a large growth in the number of visits from EUHimL language-setting users, the proportion of visits to NHSinform by users with other language settings is small and in decline. In the wider view, visits from browsers with language settings other than English is also in decline. The decision was taken in July 2017 to redouble efforts to provide Plain English content suitable for people with a reading age of 11 years and not provide machine translation on the NHSinform site. Translations of any content on NHSinform into any language is available on request. Interpreters are available on request. NHS 24 will continue to maintain an interest in Machine Translation research in meeting its equalities obligations.

8 Conclusion

In this deliverable we report on comprehensive evaluation of the translations provided in the HimL project. We have seen how our deployed and research systems have improved considerably over time, and how the quality of our translations is generally higher than a large commercial translation provider on all language pairs.

We have performed surveys of both Cochrane and NHS24 users which show that translations are useful, especially in addition to the original English, and particularly when users have low levels of English ability.

We have performed an in-depth analysis of NHS24's user needs and seen that users expect completely accurate translations on NHS24 branded web-sites. Therefore machine translated content needs to be post-edited by humans before NHS24 is willing to publish it. We have also seen, in the Cochrane post-editing study, that human translators can save considerable effort by post-editing machine translations instead of translating from scratch.

Although we have not been able to monitor web-traffic on published machine translated content for NHS24, there has been some traffic on the Cochrane machine translated text. It is anticipated that this traffic will grow with time. Our report also shows that there is a growing number of users looking at public health information, and that there is a need for multi-lingual content.

In the field of machine translation research in the last three years there has been incredibly rapid progress, and we have shown that we have not only kept up, but pushed this research forward. Sometimes this has meant that our research agenda has had to change considerably to adapt to this changing landscape and that some of the research initially envisaged in the project has not had the impact anticipated. However we have shown here that machine translation is useful to both partners to speed up post-editing. Furthermore, Cochrane is considering publishing machine translations, especially where expert translators for that language are not available.

References

Birch, Alexandra, Omri Abend, Ondrej Bojar, and Barry Haddow. 2016. "HUME: Human UCCA-based evaluation of machine translation." *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1264–1274.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. "Findings of the 2017 Conference on Machine Translation (WMT17)." *Proceedings of the Second Conference on Machine Translation*, 169–214. Copenhagen, Denmark.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurelie Neveol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. 2016. "Findings of the 2016 conference on machine translation." *Proceedings of the First Conference on Machine Translation*, 131–198. Berlin, Germany.

Huck, Matthias, Fabienne Braune, and Alexander Fraser. 2017a. "LMU Munich's Neural Machine Translation Systems for News Articles and Health Information Texts." *Proceedings of the Second Conference on Machine Translation*, 315–322. Copenhagen, Denmark.

Huck, Matthias, Simon Riess, and Alexander Fraser. 2017b. "Target-side Word Segmentation Strategies for Neural Machine Translation." *Proceedings of the Second Conference on Machine Translation*, 56–67. Copenhagen, Denmark.

Macháček, Matouš and Ondřej Bojar. 2013. "Results of the WMT13 metrics shared task." *Proceedings of the Eighth Workshop on Statistical Machine Translation*, 45–51. Sofia, Bulgaria.

Mareček, David, Ondřej Bojar, Ondřej Hübsch, Rudolf Rosa, and Dusan Varis. 2017. "Cuni experiments for wmt17 metrics task." *Proceedings of the Second Conference on Machine Translation*, 604–611. Copenhagen, Denmark.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. "Universal Dependencies v1: A Multilingual Treebank Collection." *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

Popovic, Maja. 2015. "chrF: character n-gram F-score for automatic MT evaluation." *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT-15)*, 392–395.

Rosa, Rudolf. 2014. "Depfix, a tool for automatic rule-based post-editing of SMT." *The Prague Bulletin of Mathematical Linguistics*, 102:47–56.

Sennrich, Rico, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. "The University of Edinburgh's Neural MT Systems for WMT17." *Proceedings of the Second Conference on Machine Translation*, 389–399. Copenhagen, Denmark.

Tu, Z., Y. Liu, L. Shang, X. Liu, and H. Li. 2016. "Neural Machine Translation with Reconstruction." *ArXiv e-prints*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention is all you need." *CoRR*, abs/1706.03762.

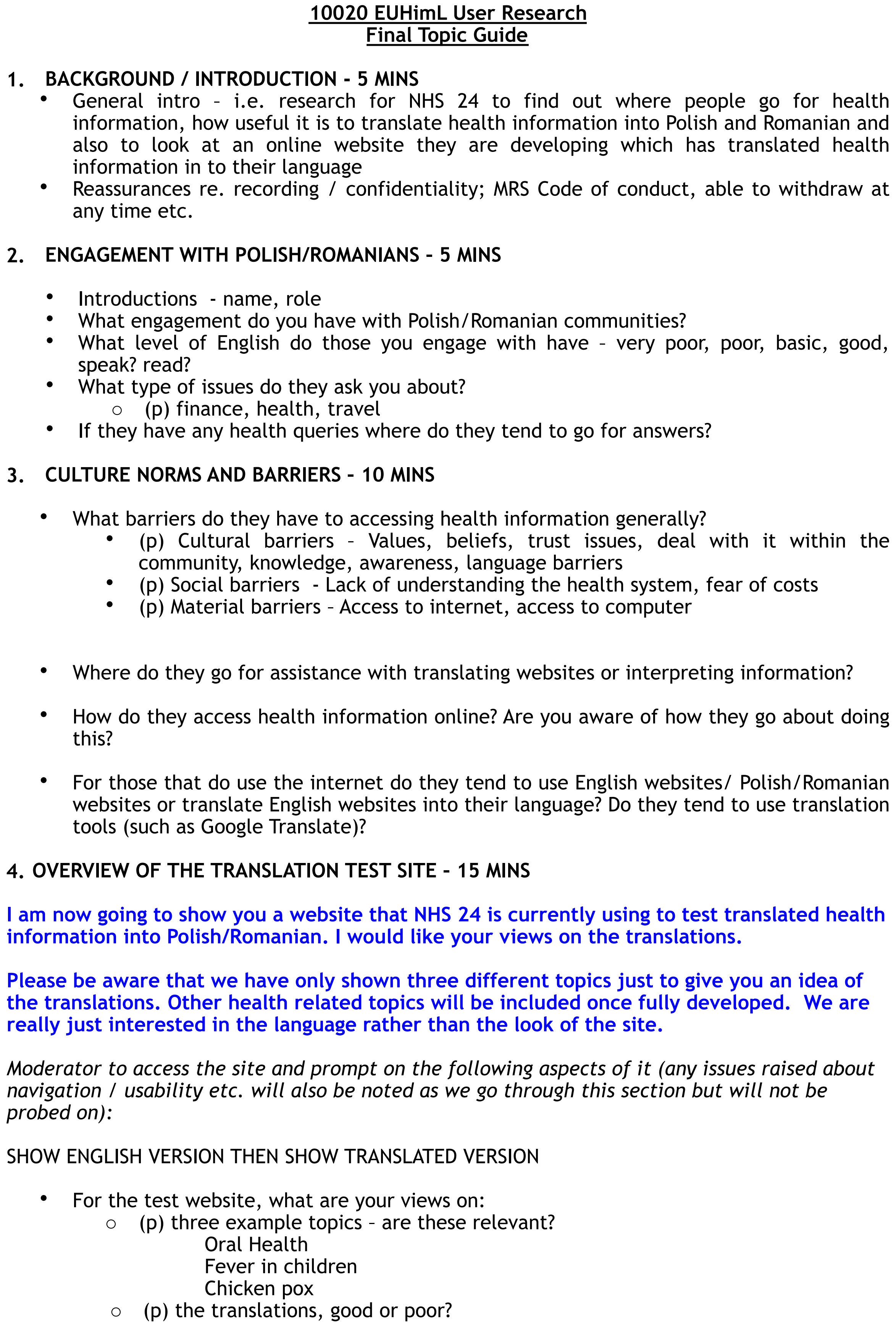
A Appendices

A.1 Detailed post-editing results

		Post-editing				Standard workflow		
	CD number	Words	Time-to-edit	Avg secs / words	PEE	Words	Time-to-edit	Avg secs / words
CS	CD001830	377	01:25:19	00:00:13	45 %	452	01:20:53	00:00:10
	CD002798	375	01:22:06	00:00:10	31 %	428	01:00:16	00:00:08
	CD007476	569	01:35:46	00:00:10	43 %	709	01:35:49	00:00:08
	CD007733	447	00:55:00	00:00:08	42 %	374	00:43:40	00:00:07
	CD007776	439	01:00:51	00:00:08	45 %	458	00:48:06	00:00:07
	CD008276	366	00:35:40	00:00:05	13 %	283	00:58:36	00:00:11
	CD009832	518	00:26:38	00:00:03	9 %	513	01:01:39	00:00:07
	CD011891	720	01:32:53	00:00:06	20 %	310	01:01:29	00:00:12
	CD012077	656	01:14:39	00:00:05	17 %	276	01:00:15	00:00:13
	CD012536	404	00:18:28	00:00:03	17 %	312	01:29:14	00:00:15
DE	CD001830	493	00:34:22	00:00:04	19 %	495	00:28:46	00:00:04
	CD002798	464	00:12:26	00:00:01	3 %	464	00:52:18	00:00:07
	CD007476	714	00:35:09	00:00:03	8 %	722	00:39:58	00:00:04
	CD007733	430	00:30:07	00:00:04	20 %	450	00:37:44	00:00:06
	CD007776	458	00:15:24	00:00:02	13 %	458	00:45:56	00:00:07
	CD008276	378	00:16:11	00:00:03	7 %	386	00:28:27	00:00:04
	CD009832	518	00:19:31	00:00:03	10 %	518	00:24:39	00:00:03
	CD011891	724	00:18:33	00:00:01	10 %	694	01:45:15	00:00:09
	CD012077	712	00:26:12	00:00:02	12 %	712	01:26:45	00:00:07
	CD012536	402	00:11:31	00:00:02	17 %	406	00:27:40	00:00:04
PL	CD001830	491	01:04:06	00:00:09	45 %	476	00:42:29	00:00:06
	CD002798	464	00:27:28	00:00:04	28 %	460	00:47:04	00:00:05
	CD007476	717	01:23:14	00:00:07	32 %	718	00:56:30	00:00:05
	CD007733	454	00:31:36	00:00:07	35 %	455	00:46:35	00:00:09
	CD007776	454	00:20:53	00:00:03	46 %	458	00:42:13	00:00:05
	CD008276	388	00:31:23	00:00:05	30 %	N/A	N/A	N/A
	CD009832	518	00:27:18	00:00:03	32 %	514	00:33:01	00:00:05
	CD011891	724	00:53:21	00:00:05	31 %	692	00:43:04	00:00:06
	CD012077	712	01:06:55	00:00:05	40 %	712	00:26:45	00:00:03
	CD012536	406	00:18:55	00:00:03	29 %	406	00:21:05	00:00:04
RO	CD001830	495	00:29:13	00:00:04	18 %	N/A	N/A	N/A
	CD002798	464	00:08:15	00:00:01	5 %	N/A	N/A	N/A
	CD007476	722	00:21:00	00:00:02	6 %	N/A	N/A	N/A
	CD007733	449	00:17:49	00:00:04	30 %	N/A	N/A	N/A
	CD007776	458	00:16:45	00:00:03	13 %	N/A	N/A	N/A
	CD008276	388	00:23:13	00:00:04	13 %	388	00:21:17	00:00:04
	CD009832	518	00:12:38	00:00:02	8 %	490	00:25:55	00:00:03
	CD011891	724	00:27:27	00:00:02	6 %	693	00:58:08	00:00:06
	CD012077	712	00:21:00	00:00:01	3 %	651	00:58:47	00:00:06
	CD012536	406	00:08:11	00:00:01	5 %	406	00:25:53	00:00:05

Table 21: Breakdown of results from post-editing pilot per PLS, task type and language. CS = Czech, DE = German, PL = Polish, RO = Romanian, PEE = Post-editing effort. N/A = The collected data was not taken into account for the analysis due to aforementioned issues with the task setup. NB: Avg secs / words was obtained by using the Avg secs / words values provided in the MateCat editing log, not by dividing Time-to-edit by Words.

A.2 NHS24 Topic Guides for first In-depth Impact Interviews



10020 EUHimL User Research
Final Topic Guide

1. **BACKGROUND / INTRODUCTION - 5 MINS**
   - General intro – i.e. research for NHS 24 to find out where people go for health information, how useful it is to translate health information into Polish and Romanian and also to look at an online website they are developing which has translated health information in to their language
   - Reassurances re. recording / confidentiality; MRS Code of conduct, able to withdraw at any time etc.

2. **ENGAGEMENT WITH POLISH/ROMANIANS - 5 MINS**

   - Introductions - name, role
   - What engagement do you have with Polish/Romanian communities?
   - What level of English do those you engage with have – very poor, poor, basic, good, speak? read?
   - What type of issues do they ask you about?
     - ○ (p) finance, health, travel
   - If they have any health queries where do they tend to go for answers?

3. **CULTURE NORMS AND BARRIERS - 10 MINS**

   - What barriers do they have to accessing health information generally?
     - (p) Cultural barriers – Values, beliefs, trust issues, deal with it within the community, knowledge, awareness, language barriers
     - (p) Social barriers - Lack of understanding the health system, fear of costs
     - (p) Material barriers – Access to internet, access to computer

   - Where do they go for assistance with translating websites or interpreting information?

   - How do they access health information online? Are you aware of how they go about doing this?

   - For those that do use the internet do they tend to use English websites/ Polish/Romanian websites or translate English websites into their language? Do they tend to use translation tools (such as Google Translate)?

4. OVERVIEW OF THE TRANSLATION TEST SITE - 15 MINS

I am now going to show you a website that NHS 24 is currently using to test translated health information into Polish/Romanian. I would like your views on the translations.

Please be aware that we have only shown three different topics just to give you an idea of the translations. Other health related topics will be included once fully developed. We are really just interested in the language rather than the look of the site.

Moderator to access the site and prompt on the following aspects of it (any issues raised about navigation / usability etc. will also be noted as we go through this section but will not be probed on):

SHOW ENGLISH VERSION THEN SHOW TRANSLATED VERSION

- For the test website, what are your views on:
  - ○ (p) three example topics - are these relevant?
    Oral Health
    Fever in children
    Chicken pox
  - ○ (p) the translations, good or poor?

- What ways, if any, will the machine translations make it easier for them to understand the health information, as compared to just having the English content?

- The translations are automatically generated and will not be perfect. Do you think that would still be preferred over just having Plain English content?

- What, if any, information is missing? Anything you would have expected that is not there? (p) is it detailed enough?, (p) does it give the right type of advice? (p) does it cover all the types of questions they may come to you with/or ask others about e.g. who to phone and when?

- What other issues might crop up (not discussed earlier) with the Polish/Romanian community using this website?
    - (p) access online, difficulty using websites, concerns about quality of translations, other language barrier issues, out-of-date, don't trust the information

- What do you think about the fact it is developed by NHS?
    - (p) How does that make you feel?
    - (p) Does that add credibility to the site? Why/why not?
    - (p) Does it add credibility to the translations? Why/Why not?

- Do you think this is something that is needed for these communities? Why? Why not?

5. PROMOTION / ACCESSIBILITY - 5 MINS

- How do you think NHS 24 can make these communities aware of a site like this?
    - Who should direct people to this kind of site?
    - What sort of information might encourage people to visit it?

- How do these communities usually find out about new and interesting websites?

6. ANY OTHER COMMENTS
Are there any other comments you want to make that you think might impact or could feed into the development of a site like this?

Thank and close

A.3 NHS24 Topic Guides for Y3 Stakeholder In-depth Impact Interviews

10020 EUHimL Extension User Research (Year 3 translations)
Topic Guide

1. BACKGROUND / INTRODUCTION - 5 MINS

- General intro - i.e. you may recall the research we did for NHS 24 that evaluated machine translations of health information on an online website . NHS 24 has developed this further and would like views on the most recent set of translations
- Reassurances re. recording / confidentiality; MRS Code of conduct, able to withdraw at any time etc.
- Introductions - name, role
- Reminder of what engagement you have with Polish/Romanian communities?

8. EVALUATION OF THE TRANSLATION TEST SITE - 25 MINS

We are now going to discuss the updated translations that NHS 24 is currently testing. We sent the link to the test website to you in advance of this discussion and I would like your views on the translations.

Please be aware that we have sent you only a few example topics to give you an idea of the translations. Other health related topics will be included once fully developed. We are just interested in the language rather than the look of the site.

CHOOSE A TOPIC, PLEASE READ THROUGH TRANSLATED VERSION FIRST

- What do you think about the translations in general? Good? Poor?
    - Why?

COMPARE TO ENGLISH VERSION THROUGHOUT DISCUSSION

- Any errors?
    - What are the errors?
        - **Incorrect word used?**
        - **Word omitted from a sentence?**
        - **Grammatical errors?**
        - **Incorrect type of word used?**
    - What do you think it should say? What is the correct word, phrase?

- Is there anything else mentioned that is unclear? Any instructions that are not clear enough?

- Are there any translations that give contradictory advice to the English version of the site? If so, what are these?

9. COMPARE THE YEAR 3 TRANSLATIONS TO YEAR 2 - 10 MINS

READ YEAR 2 TRANSLATED SITE AS A REMINDER OF THESE TRANSLATIONS

- What do you think of the latest translations compared to the previous translated site we showed you (year 2)?
    - Better/worse/no change?
    - If better, what is better?
    - If worse, how?

- The year 3 translations are generated by machine and are still not perfect. Do you think this translation would be preferred over just having Plain English content for people with little English?
    - In its current state do you think it would be used? Why? Why not?

- Do you think people with little English would use the translation alongside the English version?
    - Why/Why not?