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

Health in my Language

HimL

D5.5: Report on user surveys, impact assessment and automatic semantic metrics.

Author(s): Alexandra Birch, Ondřej Bojar, Rudolf Rosa, Juliane Ried, Hayley Hassan, Colin Davenport

Dissemination Level: Public

Date: July 31st 2017

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	July 31st 2017
Actual date of delivery	August 2nd 2017
Deliverable number	D5.5
Deliverable title	Report on user surveys, impact assessment and automatic semantic metrics.
Type	Report
Status and version	1.0
Number of pages	23
Contributing partners	UEDIN
WP leader	UEDIN
Task leader	UEDIN
Authors	Alexandra Birch, Ondřej Bojar, Rudolf Rosa, Juliane Ried, Hayley Hassan, Colin Davenport
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 Executive Summary

This deliverable appears in the Grant Agreement as *D5.5: Report on integrated semantic evaluation metric*. It was extended, with agreement from the project officer, to include the user surveys and impact studies based on the Y2 (Year 2) systems. These were due to appear in D5.4, but because of delays in the deployment of the Y2 systems the results were not available in time for this earlier deliverable. The new title reflects the enlarged scope of the current deliverable.

The body of the document begins (in Section 2) with a description of the test sets produced in this project. We have two translation test sets and also two data sets which have been annotated with the HUME human semantic annotations. Both are available online.

Section 3 provides a summary of the work done on an automatic semantic metric. We have submitted the HUME annotated data as one of the gold standard test sets for the WMT Metrics Task, and we are therefore able to determine which automatic metrics correlate closest with our semantic annotation data. The results of the metrics task will be available in September 2017 at WMT, see Bojar et al. (2017).

The rest of the deliverable is concerned with the user studies and impact assessments.

The first user study was performed by Cochrane on the benefit of machine translation as a preparatory step to human translation. It showed that post-editing machine translation speeds up translation for three of the four target languages when using the HimL Year 2 translation systems.

The second user survey was performed by Cochrane to determine if the the HimL machine translations were of a high enough standard to be useful by themselves. A sample of the Cochrane plain language summaries were translated and rated online by people who are registered as volunteers on the Cochrane website. The results showed a large improvement in acceptance of MT output between the year 2 translation systems, and the neural machine translation systems. They are obviously still not as highly rated as human translations, but were given median and mean scores of three stars out of five for three of the HimL languages. One in three people claimed that they were more useful than just seeing the original English, and this number is probably skewed against us by the large number of Cochrane users who speak fluent English.

The third user survey was performed by NHS24. This user survey had the aim of assessing the usefulness of the machine translations in the health information context and user's expectations about automatic machine translation. Although 67% of respondents reported that the translations were not accurate, this could be due to the unrealistic expectations of machine translation quality. Better translation models and better designed surveys will help to address this problem.

The final study was on the languages spoken by users of the Cochrane website. There are clearly a large number of people primarily speaking HimL languages who access the website. Publication of more HimL translated content in their languages would hopefully increase the number of page views in those languages.

2 Overview of HimL Test Sets

During the project, we created several test sets for different purposes and with different types of manual annotation. To avoid confusion, we summarize all of them in Table 1. Essentially, two types of test sets were created: (1) manually translated texts from our domains NHS24 and Cochrane, and (2) MT outputs manually annotated with HUME, i.e. our manual semantic translation quality metric. Test sets of type (1) are called "HimL Test Sets", while test sets of type (2) are called "HUME Test Sets".

The first HimL test set was described in D5.1 Test sets for HimL Languages and the second in D5.4 Report on second year's MT evaluation. The experiment which created the data in the first HUME test set was described in D5.2 Report on first year's MT evaluation and D5.3 Report on preliminary semantic evaluation metric. The experiment which created the data in the second HUME test set was described in D5.4 Report on second year's MT evaluation.

HUME Test Sets are available in github repository: https://github.com/bhaddow/hume-data HimL Test Sets are available on the project website: http://www.himl.eu/test-sets

3 Automatic Semantic Evaluation

In automatic semantic evaluation, we continued our work outlined in Deliverable 5.3, Report on preliminary semantic evaluation metric.[1] Please refer to that deliverable for the overview of our approach and motivation, as well as for human evaluation, upon which we base our automatic metrics and their evaluation.

[1] http://www.himl.eu/files/D5.3_Interim_Evaluation_Report.pdf

Name	Sentences	Source	Processing	Used for
HimL Test Set 2015	7,784	Cochrane Summaries, NHS24 web English	translated into cs, de, pl, ro	MT evaluation
HUME Test Set round 1	~339 depending on language pair	HimL Test Set 2015	manually annotated en with UCCA, machine-translated with Y1 systems (1 system per LP), manually evaluated MT with HUME	WMT16 metrics task test set, development of our automatic semantic metric (D5.3)
HimL Test Set 2017	1,511	Cochrane Summaries, NHS24 web English	translated into cs, de, pl, ro	MT evaluation, WMT17 biomedical task
HUME Test Set round 2	~1,021 depending on language pair	HimL Test Set 2015, WMT16 news test set	manually annotated en with UCCA, machine-translated with Y2 systems and two other MT systems (3 systems per language pair), manually evaluated MT with HUME	WMT17 biomedical task (MT), WMT17 metrics task (MT eval; labelled as `himltest17`), evaluation of our automatic semantic metric

Table 1: Summary of test sets created by the HimL project.

We have developed and evaluated three different methods for automatic evaluation of the machine translation (MT) quality, designed to focus on the semantic meanings of the sentences being correctly preserved:

1. **AutoDA**: A linear regression model using semantic features trained on WMT Direct Assessment scores (Bojar et al., 2016) or HUMEseg scores (Birch et al., 2016).
2. **TreeAggreg**: N-gram based metric computed over aligned syntactic structures instead of the linear representation of the translated sentences.
3. **NMTScorer**: A neural sequence classifier which assigns correct/incorrect flags to the evaluated sentence segments.

Two of the metrics, AutoDA and NMTScorer, are trainable on direct-assessment scores, while TreeAggreg is heuristical. In AutoDA and TreeAggreg, explicit dependency structures are used to provide deeper text understanding to the metrics than what is available to the usual text-based metrics; in NMTScorer, the deeper text understanding is provided implicitly by a deep neural network.

The trainable metric AutoDA, which uses deep-syntactic features, was found to perform best of all of the three proposed metrics, and achieved better correlation with humans compared to several standard metrics, such as the chrF3 metric. This metric was already described in D5.3, so we include only a brief summary of the method and its results in Section 3.1.

Section 3.2 and Section 3.3 describe the main principles of the less successful metrics, TreeAggreg and NMTScorer. All of the developed metrics have been submitted into the WMT17 Metrics Task;[2] for a more detailed description, please refer to the accompanying paper of Mareček et al. (2017).

3.1 AutoDA: Automatic Direct Assessment

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.

We developed two variants of the metric. The first one works only on Czech and uses many semantic features based on rich Czech tectogrammatical annotation (Böhmová et al., 2003). The second one uses much fewer features, however, it is language universal and needs only a dependency parsing model available.

3.1.1 AutoDA Using Czech Tectogrammatics

This metric automatically parses the Czech translation candidate and the reference translation and uses various semantic features to compute the final score. We use Treex[3] framework (Popel and Žabokrtský, 2010) to do the tagging, parsing and tectogrammatical annotation, and GIZA++ (Och and Ney, 2000) to provide word alignment.

We collect 83 various features based on matching tectogrammatical attributes computed on all nodes or a subsets defined by particular semantic part-of-speech tags. To this set of features, we add two BLEU scores (Papineni et al., 2002) computed on forms and on lemmas and two chrF3 scores (Popovic, 2015) computed on trigrams and sixgrams, so we have 87 features in total.

[2] http://www.statmt.org/wmt17/metrics-task.html
[3] http://ufal.mff.cuni.cz/treex

metric	en-cs
aligned-tnode-tlemma-exact-match	0.449
aligned-tnode-formeme-match	0.429
aligned-tnode-functor-match	0.391
aligned-tnode-sempos-match	0.416
lexrf-form-exact-match	0.372
lexrf-lemma-exact-match	0.436
BLEU on forms	0.361
BLEU on lemmas	0.395
chrF3	0.540
AutoDA (87 features)	0.625
AutoDA (selected 23 features)	**0.659**

Table 2: Selected Czech deep-syntactic features and their correlation against HUME Test Set round 1 scores. Comparison with BLEU, chrF3, and our trainable AutoDA (using chrF3 as well).

metric	en-cs	en-de	en-pl	en-ro
NIST	0.436	0.481	0.418	0.611
NIST cased	0.421	0.481	0.410	0.611
chrF1	0.505	0.497	0.428	0.608
chrF3	0.540	0.511	0.419	0.638
NIST on content lemmas	0.416	–	0.361	0.542
matching lemmas	0.431	–	0.393	0.565
matching forms	0.372	0.478	0.405	0.576
matching content lemmas	0.359	–	0.408	0.536
matching content forms	0.321	0.470	0.427	0.552
matching formemes	0.347	0.170	0.357	0.420
matching tense	-0.094	–	-0.118	0.079
matching number	0.286	–	0.205	0.404
AutoDA (linear regression)	**0.604**	**0.525**	**0.453**	**0.656**

Table 3: Pearson correlations of different sentence-level metrics on HUME Test Set round 1 dataset. Standard NIST and chrF metrics are compared with our individual features matching. AutoDA combines all the features together with the chrF3 score and the NIST score computed on content lemmas only. Other NIST scores are not included in AutoDA, since they do not bring any improvement.

We then train a linear regression model to obtain a weighted mix of features that fits best the HUME Test Set round 1 scores. The correlation coefficients are shown in Table 2, along with the individual features.

In addition to the regression using all 87 features, we also did a feature selection, in which we manually chose only 23 features with a positive impact on the overall correlation score. We see that chrF3 alone performs reasonably well (Pearson of 0.54). If we combine it with a selected subset our features, we are able to achieve the correlation of up to 0.659.

3.1.2 Language Universal AutoDA

We have found that deep-syntactic features help to train a well-performing automatic metric for Czech. Even though we have no similar tectogrammatical analysis tools for other languages so far, we try to extract similar features for them as well.

We use Universal Dependencies (UD) by Nivre et al. (2016b), a collection of treebanks in a common annotation style, where all HimL languages are present – version 1.3 covers 40 languages (Nivre et al., 2016a). For syntactic analysis, we use UDPipe by Straka et al. (2016), a tokenizer, tagger, and parser in one tool, which is trained on UD.

We distinguish content words from function ones by the POS tag. We then compute numbers of matching content word forms and matching content word lemmas. The universal annotations contains also morphological features of words: case, number, tense, etc. Therefore, we also create equivalents of tectogrammatical formemes or grammatemes.

We compute all the scores proposed in the previous section on the four HimL languages and test the correlation on HUME Test Set round 1. Similarly to Czech AutoDA, we trained a linear regression on all the features together with *chrF3* score. The results computed by 10-fold cross-validation and comparison with chrF and NIST scores is shown in Table 3.

Lang.	chrF3	TreeAggreg	Difference
en-cs	0.5403	**0.5473**	+0.0070
en-de	**0.5111**	0.5078	−0.0033
en-pl	0.4186	**0.4266**	+0.0080
en-ro	**0.6314**	0.6226	−0.0088
Average	0.5254	**0.5261**	+0.0007

Table 4: Evaluation of TreeAggreg and chrF3 baseline with Pearson's correlation to human judgments.

3.2 TreeAggreg: Tree Aggregated Evaluation

TreeAggreg is a simple sentence-level metric, inspired by HUME. 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 our work.

In TreeAggreg, we are trying to improve an existing string-based metric by applying it in a syntax-tree-based context. This is motivated by our belief that dependency trees are a good means of capturing sentence structure and semantics, preservation of which we aim to focus on. 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. Specifically, we extract the subtree spans of the sentence-root dependents; for a typical sentence structure, this means cutting the sentence into phrases corresponding to each of the arguments of the main verb We then evaluate these phrases independently with the string-based MT metric, and the resulting scores are aggregated into a final sentence-level score using a simple weighted average.

Our source code is available online.[4]

3.2.1 Evaluation

To evaluate our metric, we measured Pearson's correlation of chrF3-based TreeAggreg scores with sentence-level human judgments on the WMT16 part of the HUME Test Set round 1. For comparison, we also measure the correlation of a baseline metric, which is the vanilla sentence-level chrF3.

As shown in Table 4, our metric performs comparably to the chrF3 baseline, leading to a slight improvement for two language pairs, and a slight deterioration for the other two.

Thus, our approach of employing sentence syntactic structure into a string-based MT metric seems to affect the metric only minimally.

3.3 Neural MT Scorer

Neural MT Scorer is a model that predicts a probability for a given source/target translation pair using a simplified architecture that is based on existing NMT models with attention (Bahdanau et al., 2014). We use two LSTM encoders, one for source and one for target side. The final cell states p_s and p_t are used to measure the bilingual similarity by $\sigma(p_s^T p_t)$. The predicted number measures how much the meaning of source and target matches.

We used that model for a different task (scoring phrase table entries in PBMT) where it performed well. The training data for the model are bilingual corpus (set of sentences that should be classified as entirely correct) as well as a set of sentences that should be classified as incorrect (we obtain these by performing some random operations on the bilingual corpus).

We do not train it on data specific for the metrics task (i.e. the model is only trained to recognize correct and incorrect translations, but small differences among different translations of the same sentence might not be recognized), therefore there is a room for potential improvement.

NMTScorer only takes the input sentence and the candidate translation as its input, not requiring the reference translation. In that respect, it is more of a quality estimation system than an MT metric, resulting in poor performance when compared to other MT metrics, but making it applicable even in situations where the reference translation is not available.

3.3.1 Evaluation

We evaluated the model on the WMT16 part of the HUME Test Set round 1, but currently it performs poorly. It should be possible to improve it significantly by optimizing the training process for the metrics task (for example by adding another layer

[4] https://github.com/ufal/auto-hume/tree/rudolf

Languages	NMTScorer
en-cs	0.4099
en-de	0.3462
en-pl	0.3261
en-ro	0.4792
Average	0.3903

Table 5: Evaluation of NMTScorer with Pearson correlation to human judgments.

that uses the final representations p_s and p_t to predict human scores and fine-tune the entire model on some manually evaluated datasets). The Pearson correlation coefficients to human judgments are shown in Table 5.

3.4 Summary

We introduced three new automatic sentence-level evaluation metrics, AutoDA, TreeAggreg, and NMTScorer, which try to focus on capturing the semantic meaning of the translation. We evaluated all of them on the HUME Test Set round 1 in terms of Pearson correlation with the annotated human judgments. While two of the metrics were found to perform rather poorly, AutoDA correlates well with human judgments. It reaches Pearson coefficients between 0.45 and 0.65, surpassing other common metrics, such as BLEU, NIST, and chrF3.

4 Cochrane Post-Editing Evaluation

Accuracy is very important in the context of health information, as certain mistakes in translations could lead to patient harm. Cochrane therefore needs to carefully evaluate the output of the HimL translation systems, to gauge whether they are of an acceptable standard to publish.

One element of Cochrane's evaluation is an experiment to assess whether post-editing HimL MT is less effort and quicker than Cochrane's standard translation workflow. The aim is that little or no post-editing of translations produced by the final MT systems would be required. 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.

4.1 Introduction to Cochrane's standard translation workflow

Cochrane has extensive experience in translating its Review summaries, and they have always maintained a very high standard of accuracy. Cochrane's not-for-profit nature and limited budgets mean that there are typically no resources available to pay professional translators, and Cochrane's translation teams largely rely on volunteers. Health professionals are most likely to volunteer their time as translators, because they are often familiar with Cochrane and want to contribute to making its information available in their native language. As a result, Cochrane's volunteer translation teams mainly consist of bilingual domain experts, including for example clinicians, health researchers and medical students, and only few professional translators or people with a background in language or communications. While professional translators tend to produce more fluent translations, Cochrane has found that even professional translators specialised in health or science make significant technical errors when translating Cochrane Reviews, so a final review by a domain expert is always required. From Cochrane's perspective, in an ideal scenario, both types of experts are involved in the translation process, but domain experts are critical.

Cochrane uses a third-party translation management system (TMS), which allows translation teams to manage their translation workflow, assign content to translators, and publish translations on Cochrane's websites via an API. Like other translation software, the TMS breaks up source and target content into segments; a sentence or header is usually a segment. This facilitates translation memory (TM) matching and storage. The TMS allows translators to access Google Translate MT and Cochrane's TM while they translate, but it doesn't pre-populate segments with content from either MT or TM.

The translators participating in Cochrane's post-editing evaluation are all native speakers, familiar with Cochrane content, have a primary background in health and at least some or extensive experience as Cochrane translators.

4.2 Experiment design and setup

The EU FP7 project MateCat tool was chosen as post-editing software, because it provides a simple user interface, automatic recording of post-editing effort and time-to-edit, it allows importing personal translation memory and glossary files, and is available free of charge. Like other translation software, MateCat breaks up source and target content into segments.

To be able to compare post-editing effort and time taken to edit HimL MT with Cochrane's standard translation workflow, the same content has to be translated twice by two different translators: 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. So separate tasks have to be set up in MateCat for post-editing and standard workflow for each language and made available for the respective tasks.

MateCat typically tries to pre-populate all segments with the best match available from its global TM, which includes a large set of data contributed by the MateCat community, or, if no matches are available, with Google or Microsoft MT. Users can however add their own TM and glossaries, and prioritise those over MateCat's TM. It is also possible to disable Google or Microsoft MT.

The following setup is therefore required for Cochrane's post-editing experiment: For the post-editing task, Cochrane translates source content into the four HimL languages using the HimL MT engines, and translation memory files in TMX format are generated from the source and output content for each language. The obtained TMX files are imported into MateCat and prioritized over other available TM. As a result, all source content has a 100% match from the imported TMX, and MateCat pre-populates all segments with HimL MT from the imported TMX files, as translators open the MateCat editor to start post-editing. For the standard workflow task (labelled "human" task), manipulated "empty" TMX files are generated including the source content, but a dash instead of a translation for the target language. These empty TMX files are imported into MateCat and prioritized over other available TM. As a result, all source content has a 100% match from the imported empty TM, and MateCat pre-populates all segments with a dash, as translators open the MateCat editor. This ensures that translators need to start from scratch to mimic the standard workflow, and that they are not presented with matches from MateCat's TM or Google or Microsoft MT. In addition, existing Cochrane TM and glossaries are imported and accessible for German and Polish.

MateCat records post-editing effort and time-to-edit for each segment on a per project basis. So separate MateCat projects need to be set up for all source content per task type and language, i.e. for each source file there is a post-editing project and a human project for each language.

4.3 Post-editing pilot

A post-editing pilot was conducted in June 2017 to test MateCat and the overall experiment design for appropriateness, and to obtain initial results about the effect of post-editing HimL MT on translation efficiency.

Cochrane selected three Plain Language Summaries (PLS) that had not been previously translated into German nor Polish by Cochrane's volunteer teams to avoid bias through potential translation memory matches. The three PLS were translated into the four HimL languages using the HimL Y2 neural MT engines. Translation memory files in TMX format were generated from the source and output content for each language, and empty TMX files were also generated for the three PLS as described above.

Separate projects were set up for each PLS per task type and language, i.e. for each PLS there was a post-editing project and a human project for each language, so a total of 24 projects. All projects were labeled accordingly with PLS identifier (CD number), language identifier, and task type, e.g. CD003650-CS-post-editing.

Two translators per language participated in the pilot. The hyperlinks to the MateCat projects were distributed to the translators, and 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 human task for the same PLS. Cochrane also ran a virtual training session for the translators on how to use MateCat. The translators had four weeks to complete the task.

4.3.1 Results from the post-editing pilot

The editing log of each project was exported from MateCat in full and data was compiled into one spreadsheet. The analysis focused on time-to-edit, average seconds spent on editing per word, and post-editing effort (PE Effort). The totals and averages for those data were calculated by task type and language and are presented in Figure 1. A full break-down of results per PLS, task type and language is available in Figure 2.

Post-editing was overall quicker than human translation for Czech, German and Romanian, both in terms of total time-to-edit and average seconds spent on editing per word. For German in particular the speed-up was significant: On average, post-editing took less than half the time as human translation. However, for Polish, human translation remained slightly quicker than post-editing both in terms of total time spent on editing and average seconds per word.

The average PE Effort for the post-editing task was also on a similar level for Czech, German and Romanian: between 18-21% of MT had to be post-edited. Polish, on the other hand, registered a PE Effort of almost 33% for the post-editing task, which reflects the longer editing times that were recorded for Polish.

These results were echoed by informal feedback from translators which was collated during a virtual debrief session. For each language, translators had decided to divide up the task types 1:2 between them, i.e. each translator worked on two human

	Czech	German	Polish	Romanian
Post-editing				
Words	1270	981	1324	1354
Total time-to-edit (hh:mm:ss)	02:19:42	02:13:22	02:27:28	01:11:42
Avg secs/word	6.5	6.0	6.9	3.7
Avg PEE	20.71%	20.54%	32.60%	17.91%
Human				
Words	1081	1049	1260	1311
Total time-to-edit (hh:mm:ss)	02:45:03	03:11:35	02:21:55	01:35:16
Avg secs/word	10.6	13.5	6.7	5.5
Avg PEE	88.97%	91.19%	94.78%	93.25%

Figure 1: Summary of results from post-editing pilot by language and task type. PPE = Post-editing effort.

and one post-editing task, or the other way around. So each translator worked on each task type. The Czech, German and Romanian translators reported that they found HimL MT was of surprisingly good quality, and post-editing was clearly quicker than translation from scratch. The Polish team though found that HimL MT was of poor quality and post-editing took longer than translation from scratch.

4.3.2 Limitations of the collected data

The detailed editing log included several segments across all languages with time-to-edit records that appeared to be unreasonably long, in some cases suggesting translators took up to several hours for a single segment. After consultation with MateCat support staff, it became clear that the recorded times most likely included breaks that translators had taken while working on the tasks, because the timer doesn't stop when translators leave the MateCat editor open in their browser, and work on other tasks. Translators also confirmed during the debrief that they did get interrupted in between tasks and, in some cases, took a break before completing a project.

MateCat further explained that the editing log is supposed to exclude any segments that are translated "too fast" (in less than 0.5 seconds per word) or that "take too long" to translate (over 25 seconds per word). However, the exported data included segments that took too long and also some that were translated too fast according to MateCat's definition. This was reported to MateCat and they were working on resolving this bug.

The extremely long and some extremely short time-to-edit records meant that the overall data was somewhat skewed. To be able to analyze the data and gain at least rough insights, Cochrane applied the MateCat definitions manually and removed any segments with editing times longer than 25 seconds per word, or shorter than 0.5 seconds per word from the log. This however means that the number of words varies between different languages and task types, and is not entirely the same in terms of size and content.

Translators have been provided with clear instructions on how to avoid the same issue in the upcoming post-editing evaluation.

4.4 Conclusion

The results from the post-editing pilot provide a baseline to compare Y3 HimL MT post-editing to.

Despite the small content sample and reported limitations, the results from the post-editing pilot have been encouraging with post-editing of Y2 HimL neural MT outperforming Cochrane's standard translation workflow for three of the four HimL languages in terms of time needed for editing, and positive translator feedback. The less positive results and feedback for Polish

		Post-editing				Human			
	CD number	Words	Time-to-edit	Avg secs / words	PEE	Words	Time-to-edit	Avg secs / words	PEE
CS	CD003650	440	00:31:13	6.0	21.60%	335	01:07:33	11.2	100.00 %
	CD011894	342	00:46:09	6.8	24.45%	147	00:40:13	14.2	72.22%
	CD012499	488	01:02:20	6.7	16.07%	599	00:57:17	6.4	94.70%
DE	CD003650	415	01:12:29	9.8	17.75%	164	01:03:54	22.7	100.00 %
	CD011894	5	00:00:11	2.2	33.33%	279	01:02:05	11.4	81.25%
	CD012499	561	01:00:42	6.0	10.54%	606	01:05:36	6.5	92.32%
PL	CD003650	396	01:00:11	9.1	37.40%	346	00:44:27	7.6	100.00 %
	CD011894	342	00:28:39	5.6	24.05%	346	00:40:57	7.1	85.71%
	CD012499	586	00:58:38	5.9	36.35%	568	00:56:31	5.5	98.63%
RO	CD003650	402	00:21:13	3.7	22.09%	384	00:22:33	4.5	97.59%
	CD011894	346	00:20:35	3.6	16.95%	322	00:35:23	7.8	86.17%
	CD012499	606	00:29:54	3.8	14.68%	605	00:37:20	4.3	96.00%

Figure 2: 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.

HimL MT and post-editing are in line with previous findings from the Y2 ranking evaluation whereby Y2 HimL neural MT was not performing as well for Polish as it did for the other HimL languages.

4.5 Outlook

The post-editing experiment will be re-run when the final HimL Y3 systems are available, and sample content will be expanded to include up to 10 PLS. Cochrane will attempt to avoid the issues related to the recording of time-to-edit that occurred in the post-editing pilot by carefully instructing translators on best editing practice. The aim is that a larger dataset, optimized HimL MT engines, and improved data collection will confirm and improve on the promising results from the pilot.

5 Cochrane User Survey Evaluation

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.

5.1 Survey design and display

Cochrane randomly selected 20 Plain Language Summaries (PLS) and translated them into the four HimL languages, Czech, German, Romanian, and Polish, using the HimL machine translation systems as described in D4.2/5 (4.1.2 Publication and Display of Translations on Cochrane Website).

The translations were published on cochrane.org as part of the four dedicated language versions. Users had access to the English original via the language toggles on top of the window. Cochrane managed user expectations by displaying a message explaining that a machine translation engine produced the translations as part of the HimL project, and providing a link to further information and a contact email. As users accessed the translated PLSs, after a few seconds, a survey pop-up in their language dropped down from the top of the browser and asked users to rate how easy the translation was to understand. The posed question was:
"The translation below was generated using machine translation software. How easy is it to understand?"

Below the question, there were five empty stars for people to assign a rating out of five. An explanation of the value of the stars appeared when users hovered their mouse: from 1 star for very hard, 2 stars for hard, 3 stars for neutral, 4 stars for easy,

to 5 stars for very easy. Questions and ratings were translated into the four languages. A Czech example can be viewed here: http://www.cochrane.org/cs/CD009678/ or in Figure 3.

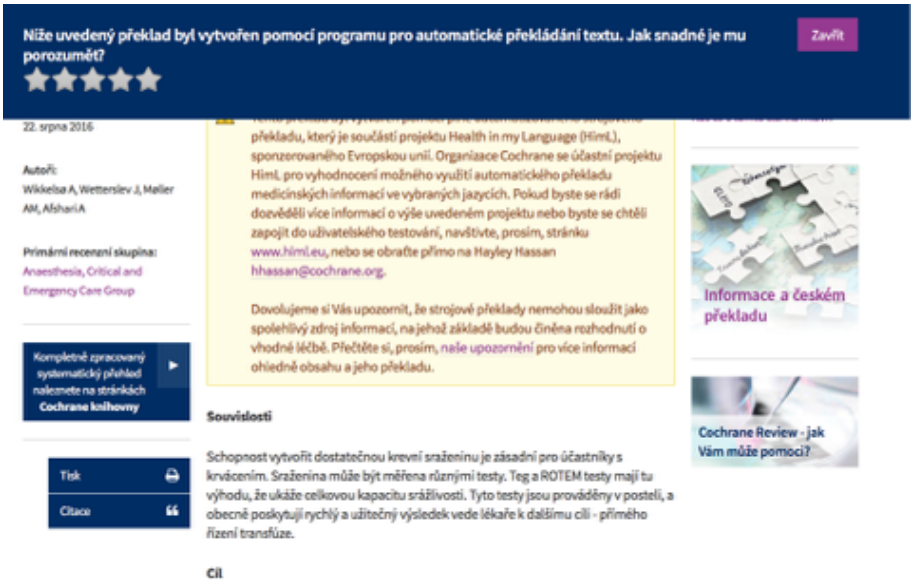


Figure 3: Czech PLS translation and survey pop-up on cochrane.org.

To collect comparison data, Cochrane added a similar pop-up survey to 20 randomly selected PLS that had previously been translated into German and Polish by Cochrane's volunteer teams. The posed question was: "The translation below was produced by a volunteer translation team. How easy is it to understand?" As above, users were asked to rate the ease of understanding out of five stars. A Polish example is available here: http://www.cochrane.org/pl/CD010070/.

5.2 Rationale for the survey design

After considering a longer survey, and internal discussions with experts in survey design and qualitative research around Cochrane content, Cochrane considered that the above described survey design would yield the most useful results.

Firstly, respondents are more likely to spend time answering a quick pop-up question than following a link to a fuller survey, so the response rate was expected to be much higher for a pop-up survey than a longer, off-site survey.

A longer survey would have given Cochrane the chance to elicit user information (whether users have a health background, their level of English, etc.), and ask more questions about the usability and quality of the translations, however, it was felt that these answers would not have helped draw stronger conclusions regarding how easy the machine translations were to understand, or their usefulness. These types of questions are open for interpretation by the respondent and potentially biased for different reasons (e.g. the English original may be of bad quality, the user doesn't like a particular research result, people have subjective expectations and judgment about quality and usefulness of texts in general). These limitations could probably only be addressed better by qualitative focus groups or interviews, which Cochrane doesn't have the resources for at this point.

The comparison survey ran on translations produced through Cochrane's regular translation process, i.e. by its volunteer translation teams, to contextualize the rating of the machine translations. Having a comparison let Cochrane determine if HimL machine translation is better, worse or comparable to its volunteer translations. The results of this comparison may influence Cochrane's future translation strategy decisions.

The setup also made it possible to ensure that responses were collected as direct feedback on specific PLS translations, which in turn allowed to detect whether certain PLS receive particularly bad or good results, which may be linked to the quality or content of the PLS, not just the translation itself.

Finally, by having the survey pop-up on different PLS in the four languages, and having the PLS indexed by search engines, the potential audience for the survey could be widened, see also Section 5.3 below.

5.3 Survey promotion

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 language)

and its Community website.

In addition, the survey was posted as a task for each language on TaskExchange, an open platform for people to post tasks related to Cochrane work to find people with the relevant skill sets to help them, usually on a volunteer basis.

The machine translations were displayed on cochrane.org in the same way as Cochrane's volunteer translations, and users could find them by searching and filtering in their language. In addition, they were indexed to appear in search engine results. This was important to widen the audience since most users find cochrane.org via Google. In 2016, Google searches accounted for 67% of all cochrane.org traffic.

5.4 Results of Y2 MT evaluation

The user survey was run on translations produced by the Y2 HimL MT and responses were collected during the period 1 March to 31 May 2017.

5.4.1 Y2 machine translation survey results

A total of 469 responses were received for the machine translation survey across the four languages. A full breakdown of the results from the machine translation survey is available in Appendix A, Figure 6.

In Table 6, three different values were calculated: the mean, median and mode. The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. The median indicates the middle value out of all scores when arranged numerically. The mode indicates the star rating that was given most often.

Language	No. Responses	Mean	Median	Mode
Czech	103	2.2	2	2
German	147	2.1	2	2
Polish	99	1.9	2	1
Romanian	120	1.9	1	1
All	469	2.0	2	1

Table 6: Y2 machine translation survey results: Mean, median and mode star ratings by language. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy.

The median and mean rating for all languages was 2 out of 5 stars, while the mode was 1. Czech and German received a slightly higher mean rating than Polish and Romanian. Similarly, the most commonly given scores were 2 out of 5 stars for Czech and German, but only 1 out of 5 for Polish and Romanian.

Looking at ratings of individual PLS (Appendix A, Figure 6), 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 CD000501 received ratings of 1, 2, 3, 4 and 5 stars. Generally, ratings ranging from 1 to 3 stars for the same PLS were not uncommon. Some Romanian and German machine translations received 5 star ratings, although very few compared to lower ratings, while the highest Czech and Polish scores were 4 stars.

Certain PLS were perceived as easier to understand than others, for example, CD003566 received an average rating of 3.6 out of 5, while CD011319 received an average rating of 1.7 out of 5. CD003566 is a very short PLS describing a simple intervention, while CD011319 is much longer, and describing a more complex health problem, which may have affected those ratings.

5.4.2 Volunteer translation survey results

A total of 181 responses were received for the volunteer translation survey across the two languages. A full breakdown of the results from the volunteer translation survey is available in Appendix A, Figure 7.

Table 7 shows that the translations produced by Cochrane's Polish and German volunteers yielded very similar scores, with the average rating falling between easy and very easy, and both languages achieving a median and mode of 5 out of 5 stars.

Looking at ratings of individual PLS (Appendix A, Figure 7), there was less variation between different PLS and different language translations of the same PLS, and less wide-ranging ratings for the same PLS than for the machine translations. Overall, the volunteer translations received more consistent scoring than the machine translations. Eight PLS received only 4- and 5- star ratings, and 5 stars was the most common rating for each PLS, except CD006941, which had the same number of 4- and 5-star ratings. There was only one 1-star and one 2-star rating respectively.

Language	No. Responses	Mean	Median	Mode
German	69	4.5	5	5
Polish	112	4.6	5	5
All	181	4.6	5	5

Table 7: Volunteer translation survey results: Mean, median and mode star ratings by language. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy.

5.4.3 Conclusions from Y2 MT user evaluation

As outlined above, Cochrane's volunteer translations outperformed the HimL machine translations in terms of how easy they were to understand for German and Polish users, and comprehension ratings provided by users for PLS translated using Y2 HimL MT engines were overall at the lower end. While the results suggest that the Y2 systems do not provide an appropriate level of comprehension for Cochrane audiences, the data provide a baseline to compare Y3 system user evaluation to.

Following discussions with the HimL consortium members as to whether the posed question was sufficient to gain an understanding of the usefulness of the machine translations compared to not having a translation at all, and Cochrane decided to adapt the survey as explained in Section 5.5 below.

5.5 Y2 neural MT user evaluation and survey adaption

In a second iteration, the survey is currently being re-run on the same set of PLS since 12 June 2017, but translated using the latest available neural HimL MT systems. The survey has been reset to start collecting responses from zero, and promotion has also re-started. This allows Cochrane to user test whether the neural MT models are likely to outperform the phrase-based Y2 models in the four languages, and will feed into decisions on the final system releases for the HimL project. In addition, in this second iteration, Cochrane added a second question to the user survey to test whether this could provide additional insights into the usefulness of the translations:

"Is this translation more useful to you than only seeing the original English text?"

The aim is to elicit whether users prefer to read health information on cochrane.org in their own language, despite it not being perfect perhaps, instead of only having the original text in English.

The comparison survey on German and Polish volunteer translations is also being re-run with the new question added.

5.5.1 Preliminary results of Y2 neural MT user evaluation

A total of 389 responses were received for the machine translation survey across the four languages from 12 June to 11 July 2017. A breakdown of results by PLS and language is available in Appendix A, Figure 8 and Figure 9.

As before, the mean, median and mode were calculated for each language and all languages combined.

Language	No. Responses	Mean	Median	Mode
Czech	15	3.2	3	3
German	109	3.1	3	3
Polish	51	2.6	3	3
Romanian	214	1.6	1	1
All	389	2.2	2	1

Table 8: Y2 neural machine translation survey results: Mean, median and mode star ratings by language. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy.

Looking at Table 8 and compared to the first survey, the mean, median and mode for German and Czech improved from 2 to 3 stars. For Polish, the mean improved from 1.9 to 2.6, the median went up from 2 to 3 stars, and the mode from 1 to 3 stars.

The Romanian results are however slightly worse compared to the first survey, with the mean down from 1.9 to 1.6, and mean and mode remaining at 1 star. User feedback has been received about the Romanian machine translations indicating that random sentences may have been introduced into the translations that do not match the source. HimL system developers checked and confirmed that this was an issue produced by the neural MT engine. This problem may explain why Romanian scores are low compared to the other languages.

Language	Yes, more useful	No, not more useful	Question Skipped
Czech	11	4	0
German	61	45	3
Polish	0	48	3
Romanian	36	125	53
All	108	222	59

Table 9: Y2 neural machine translation survey results: Responses by language whether the machine translation was more useful than only seeing the English text.

Table 9 gives an overview of the responses to the second survey question: "Is this translation more useful to you than only seeing the original English text?" For Czech and German, more respondents answered "yes" than "no". Almost 60% of German respondents said "yes", and more than 70% of Czech respondents said "yes", although there were only a few responses overall for Czech. For Polish and Romanian, however, a clear majority answered "no". For Romanian, only 16% of respondents answered "yes", 25% skipped the question, and 58% said "no", a result that may again reflect the issue detected about the Romanian machine translations. For Polish, there was not a single "yes" response, which is surprising given that the Polish star ratings were almost on the same level as for German and Czech. An explanation could be that the respondents may mostly have had good English skills, but since the survey was anonymous and didn't ask for users' English skills, this cannot be confirmed.

Looking at results of individual PLS (Appendix A, Figure 8 and Figure 9), there was again variation between different PLS and different language translations of the same PLS. Since Czech, German and Polish ratings had improved, but Romanian remained low, there were even wider ranges of ratings for the same PLS and most PLS had ratings ranging from 1 to 4, or 1 to 5 stars. While Czech, German and Polish ratings were more equally distributed between 1 to 5 stars with a small peak at 3 stars, most Romanian ratings were clearly 1 star and 2 stars.

For the most part, for the second question on whether the machine translation was more useful than only seeing the English text, results of individual PLS were fairly consistent within, but not across languages. For Czech and German, most PLS received a mix of "yes" and "no" answers, and often received more "yes" than "no" answers. For Romanian, most PLS received "no" answers, and either less or no "yes" answers. Almost all Polish PLS received only "no" answers, except for three skipped answers.

5.5.2 Volunteer translation survey results

A total of 64 responses were received for the volunteer translation survey across the two languages. A full breakdown of the results from the volunteer translation survey is available in Appendix A, Figure 10 and Figure 11.

Language	No. Responses	Mean	Median	Mode
German	40	4.6	5	5
Polish	24	4.8	5	5
All	64	4.6	5	5

Table 10: Volunteer translation survey results, second survey: Mean, median and mode star ratings by language. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy.

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, and the median and the mode being 5 for both languages.

Language	Yes, more useful	No, not more useful	Question Skipped
German	17	21	2
Polish	4	18	2
All	21	39	4

Table 11: Volunteer translation survey results, second survey: Responses by language whether the translation was more useful than only seeing the English text.

For the second survey question, Table 11 shows that 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", but still with

a slight majority of "no" answers. These results do not seem to reflect the high comprehension scores awarded to the volunteer translations in question one. It could be that a majority of the respondents had a very high level of English, and therefore didn't feel they needed a translation in their language, especially if it is not perfect. The number of responses available for these preliminary results is relatively low though, so perhaps a larger dataset would also change the observed pattern.

Results of individual PLS (Appendix A, Figure 8 and Figure 9) were quite consistent between different PLS and different language translations of the same PLS. Six PLS received only 5-star ratings, and 13 PLS only 4- and 5- star ratings. Only one 3-star and one 2-star rating was given respectively, and no 1-star rating at all.

For the most part, for the second question on whether the volunteer translation was more useful than only seeing the English text, results of individual PLS were relatively consistent. For German, about half of the PLS received a "no" answer, or if it was a mix of "yes" and "no" answers, then mostly more "no" than "yes" answers. For Polish the majority of PLS received only "no" answers. The relatively low turnout was distributed fairly evenly across PLS, with each receiving between 2 to 6 replies, and there weren't any strong results one way of the other for any PLS.

5.5.3 Conclusions

The preliminary results from the second survey suggest that there has been an improvement in the quality of German and Czech translations produced using the Y2 neural MT systems compared to the Y2 phrase-based MT models. Cochrane's volunteer translations still outperformed the HimL machine translations in terms of how easy they were to understand for German and Polish users, but comprehension ratings provided by users for PLS translated using Y2 HimL neural MT engines improved for Czech, German and Polish. While the results for Romanian were not as promising, the survey helped detect a problem with the Romanian MT engine, which explains the lower scores awarded to Romanian in comparison to the other HimL languages.

The second survey confirmed the high quality of the German and Polish volunteer translations, however, the additional question that was supposed to provide more insight into the usefulness of translations, seemed to suggest that the volunteer translations were mostly not useful despite their high-quality ratings, while at least Czech and German machine translations were judged useful more often than not despite only moderate comprehension ratings. These results could potentially have been affected by different English skills of respondents, and the relatively small sample of responses for volunteer translations.

5.6 Outlook

Cochrane will continue the second survey to gather more responses until the final Y3 HimL MT systems are deployed, and will then evaluate whether the patterns observed on the preliminary results persist. The final user survey will then be re-run using Y3 HimL MT systems and results will be compared with previous surveys. The aim is that a larger dataset and optimized HimL MT engines will confirm and improve on the promising results from the second survey.

6 NHS24 User Survey

The user survey is part of the EUHimL Workpackage 5 Evaluation, Task 5.4 User Acceptance Testing and contribution to Deliverable 5.4 report on Second Year's MT Evaluation. Content for user acceptance testing was a small sub-set of NHSinform redeveloped content, chosen to be of relevance to the target survey population.

Users were approached via community with whom NHS 24 has an ongoing relationship.

6.1 Objectives

The user survey had the aim of assessing the usefulness of the machine translations in the health information context and user's expectations about automatic machine translation. Objectives for the survey were:

- Determine if the internet is used to access health information and how the information is found
- Does the survey website contain useful health information
- Is the translation accurate?
- Are any of the words or phrases used wrong or inappropriate in the health context?
- Is it useful to have English alongside your own your language?

6.2 Approach

The NHS 24 HimL test site contained examples of translated content of relevance to the target audience in English, Polish and Romanian. Suitability of content was determined by frequency of visits to the NHSinform website from browsers with test language settings and discussion of health topics of interest with community coordinators.

A number of third and public sector organisations who support the Polish and Romanian communities were contacted, as well as schools and universities with a large number of Polish and Romanian students. We then set out to explain to the organisations what HimL is and what the benefits are. Service users were then encouranged to contact us. We then either instructed or met with participants to carry out exercise sessions.

A 10 question survey (Appendix B) was developed on Survey Monkey and consisted of 2 parts. The first part asks about how people access and use of health information and NHS services. The second part included links to test website with health information from www.nhsinform.scot translated using the Y2 HimL translation engine. Respondents were asked to access the site and feed back on the usefulness and accuracy of the translations.

The survey was available to subjects from a range of third sector and public sector organisations who support the Polish and Romanian communities including Glasgow Health and Social Care Partnership, Renfrewshire Polish Association, NHS Scotland Territorial Health Boards during April 2017 .

6.3 Results

The survey was closed on 24/05/17. When the survey closed, the Polish User Survey had been completed 26 times. The Romanian User Survey has had 0 completions. The results of the Polish user Survey are generated from SurveyMonkey (Appendix B) and summarised below.

6.3.1 How people access health information

58% of 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. Almost 80% of searches were of NHS websites; 45% of Google.

6.3.2 How people use NHS services in Scotland

50% of respondents reported that they had difficulty accessing health information in their own language. There was an even split of 27% of people who used friends/family to translate for them and those who used a translator or interpreter. 82% reported using Google Translate to access health information. None of the respondents reported using a telephone interpretation service. 56% of people reported that having access to translated content would not have stopped them using the service. Reasons included wanting to see a doctor if something was worrying them and needing medication.

89% of respondents reported that the website did contain information which was useful to them or to friends/family living in Scotland. It is not clear, however, if they refer to the translated information or the source information which is written in English.

6.3.3 Translations

The results of this feedback show that 67% of respondents reported that the translations were not accurate (See slide 9/11 in Appendix B). This result is disappointing but we argue that it says more about the design of the survey than about the usefulness of the machine translation output.

6.4 Example Errors

We record some interesting comments given by participants in the experiment:

1. **Using words out of context**
    - “pasożytować dziecka z zimną wodę” - I think you meant to discourage sponging with cold water, but pasożytować means “being a parasite”. There are a few similar examples, but in general the text is readable and most infomation is accurate.
    - “Sometimes there are a few incorrect words/phrases used in the wrong context which makes the website harder to understand.”

- “mleko matki lub wzoru” - wzoru/wzór means mathematical equation, but here should read "mleko modyfikowane" i.e. formula milk
- The translation is very very broken. It would be some what understandable from the context, however it really is not grammatically correct. It is not just particular words... The translator must now be working properly. Most of verbs are correct, names and headlines. But complex sentences are out of order.
- Incorrect spelling, incorrect punctuation, words and phrases used in the wrong context

2. **Language which is not used by Polish speakers anymore**

   - I cannot give you the exact example, but I remember many instances especially with instructions on how to correctly use equipment or medication or exercise, there’s been multiple words used wrongly. This made it more difficult and confusing. On some occasions you also use words that are not used by Polish speakers on daily basis. Again, very confusing.
   - many sentences doesn’t make sense, are confusing or written in the language nobody’s using anymore.

3. **Poor English**

   - Spelling mistakes in the English version! (language??) The words cannot be translated into other languages if the original words are misspelled...

6.5 Outlook

Machine translation will not be completely accurate for the foreseeable future, but it is hard to argue that it is not useful given that it is used by millions of people on a daily basis and that 82% of the respondents already use Google Translate to access health information. For the next user survey we will apply stronger translation models and we will also improve the design of our user survey.

7 Impact Study: Languages and Translation

7.1 Background to publication of Cochrane translations and its effect on access to Cochrane information

The website cochrane.org houses the Abstract and Plain Language Summary (PLS) sections of all original English Cochrane Reviews, and has already been partly translated into 14 languages by Cochrane’s community of translators. Cochrane’s translation teams largely rely on volunteers, and some projects have run longer than others, so only a subset of the more than 7200 Cochrane Reviews has been translated into each language so far. As of June 2017, more than 20,400 translations have been published on cochrane.org, but numbers vary significantly between different languages (Table 12).

Of the four HimL languages, German and Polish volunteer translation projects were established before the start of the HimL project, i.e. Cochrane has been publishing human translations in those languages since 2014 and 2015 respectively. In addition, as part of the HimL project, cochrane.org has been translated into Czech and Romanian, which allows hosting the HimL machine translations of Cochrane PLS for these languages alongside existing human translations in other languages.

Cochrane has been using Google Analytics to monitor usage of its website. They consider the most meaningful information available from Google Analytics to analyse the effect of translations on access by users speaking different languages to be data about the number of visits from specific countries, as well as the language users have their browser set to. Given that not all Internet users set their browser to their native language, and that the location of users cannot always be identified reliably, these measures are not perfect, but do still give a good overall picture and allow to monitor trends over time. So Cochrane has been collecting this information for the different languages for several years.

In the past, the publication of translations in different languages has shown a tremendous effect on access to cochrane.org, which clearly demonstrates the need for translations. In 2016, 66% of all visits to cochrane.org were made using web browsers set to a non-English language[5]. This is a stark contrast to 2012, before Cochrane published translations on its website, when 68% of visits came via an English browser and the top four countries accessing cochrane.org were the USA, UK, Australia and Canada. Countries speaking the languages that are published on cochrane.org now dominate the top 20 (Table 13), particularly Spanish- and French-speaking countries which make up a big part of the audience. This not only reflects the large populations

[5] "Translation Annual Report 2016." Cochrane Community. 2017, http://community.cochrane.org/sites/default/files/uploads/inline-files/Translations%20Annual%20report%202016.pdf

Language	Number of published translations
English original Review	7,224
Croatian	2,296
French	4,847
German	1,138
Japanese	1,199
Korean	48
Malay	568
Polish	531
Portuguese	594
Russian	181
Simplified Chinese	206
Spanish	6,858
Tamil	633
Thai	24
Traditional Chinese	296

Table 12: Number of published translations of Cochrane Abstracts or PLS as of June 2017.

speaking those languages (over 272 million French speakers and over 570 million Spanish speakers worldwide[6]), but also the high number of available translations in those two languages – 95% of all Cochrane Reviews have a translation in Spanish, 67% in French.

1.	United States	11.	Peru
2.	Mexico	12.	Brazil
3.	France	13.	India
4.	Spain	14.	Venezuela
5.	United Kingdom	15.	Ecuador
6.	Argentina	16.	Croatia
7.	Colombia	17.	Germany
8.	Canada	18.	Japan
9.	Chile	19.	Belgium
10.	Australia	20.	Russia

Table 13: Top 20 countries visiting cochrane.org in 2016.

A less obvious example is Croatian: With a much smaller language community (about 21 million native speakers across ex-Yugoslavia[7]), but more than 2,300 published translations, Croatia ranked 16th in the list of countries accessing cochrane.org in 2016 - in 2012, Croatia was ranked 54th. In this case, the small amount of health information available in Croatian on the Internet likely contributes to the success of the Croatian Cochrane translations.

Most traffic on cochrane.org comes via Google search engines – in 2016, Google searches accounted for 67% of all cochrane.org traffic . People largely find and access Cochrane information, because they search for health information in their language online, not primarily because they search specifically for Cochrane information. In addition, extensive dissemination efforts from Cochrane's translation teams via social media, newsletters, partnerships and press activities contribute substantially to translation access. For example, in 2016, 20% of visits to cochrane.org from Brazil came via Facebook, which is used to promote Portuguese translations on a regular basis.

7.2 Analysis of web traffic related to HimL languages and countries

For the HimL project, Cochrane has been collecting access data for its website specific to the languages and countries relevant to the project in order to monitor the effect of translations in the HimL languages on access by users speaking those languages.

[6] "List of languages by total number of speakers." Wikipedia https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers#Estimates_by_language.

[7] "Serbo-Croatian." Wikipedia https://en.wikipedia.org/wiki/Serbo-Croatian

7.2.1 Visits by HimL countries – comparison of January-June 2016 with January-June 2017

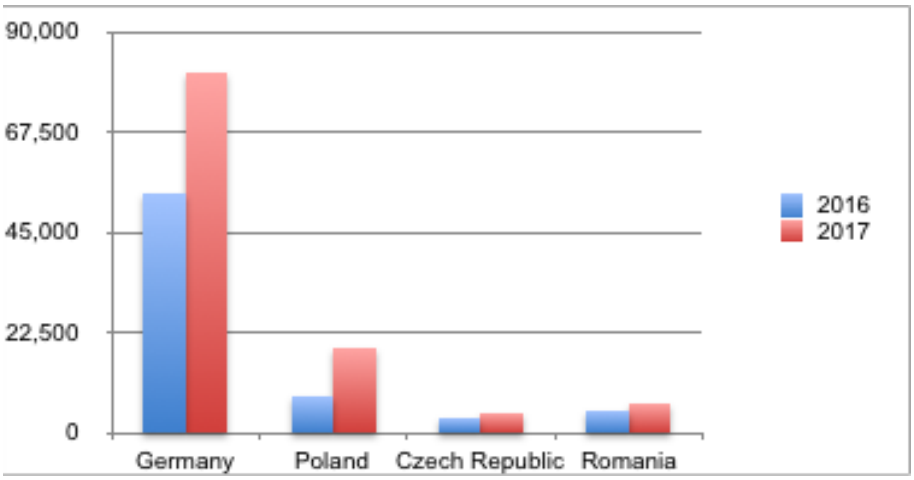


Figure 4: Number of visits to cochrane.org from users of different countries Jan-Jun 2016 compared to Jan-June 2017.

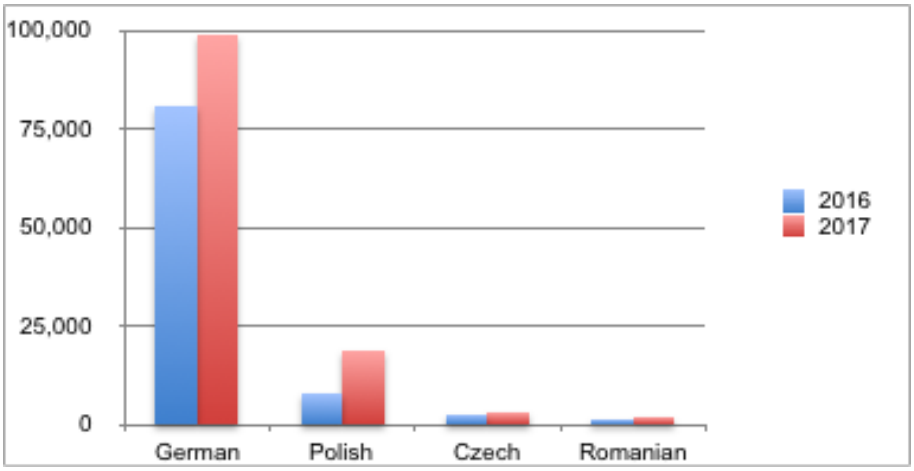


Figure 5: Number of visits to cochrane.org from users of browsers set to the HimL languages Jan-Jun 2016 compared to Jan-June 2017.

Visits to cochrane.org by users in Germany, Poland, Czech Republic and Romania have all increased comparing the period January to June 2016 to the same 6-month period in 2017 (Figure 4). Since Cochrane have been publishing German and Polish human translations for some time, Germany and Poland have been well ahead of the Czech Republic and Romania. Continuous addition of new translations combined with dissemination activities that have been built up for some time, have facilitated a substantial increase of visits by users from Poland (up 133%) and Germany (up 51%). Nonetheless, the addition of a Czech and Romanian version of cochrane.org, the publication of selected HimL machine translations and the promotion of the user survey in those languages in the first half of 2017 have also led to an increase in visits by users from the Czech Republic (up 21%) and Romania (up 36%).

7.2.2 Access by visitors using browsers set to HimL languages

Similarly, visits to cochrane.org by users with web browser languages set to German, Polish, Czech and Romanian have all increased comparing the period January to June 2016 to the same 6-month period in 2017 (Figure 5). Visits by German-language browsers are up 23%, Polish access is up 132%, Romanian is up 49%, and Czech, while only registering a small increase, is still up by 6%.

7.3 Outlook

Cochrane will continue to monitor traffic on its website specific to the HimL languages for the duration of the project to evaluate the effect of translations on access by speakers of the HimL languages. The expectation is that by adding more HimL machine translations to the site and inviting feedback via the user survey again for the Y3 systems, and by continuing the dissemination efforts in the HimL languages, it will be possible to demonstrate that more users speaking those languages are accessing the available health information because they can find what they are looking for in their language.

8 Conclusion

In this deliverable we have described the test sets that we have developed in the HimL project. We have discussed the work done on developing automated semantic translation metrics and we have described user evaluation and impact studies performed at Cochrane and at NHS24.

In the next six months we will perform one final ranking evaluation, comparing the latest research models with the HimL Year 2 and 3 systems and with Google. We will perform further user studies with NHS24 and Cochrane. Finally we will report on website statistics for users across HimL languages and other languages of interest.

Appendices

A Results Cochrane User Survey

CD number (unique ID of Cochrane Reviews)	Language	Number of 5 star ratings	4 stars	3 stars	2 stars	1 star
CD000501	Czech				2	2
	German			3	2	4
	Polish		1		1	1
	Romanian	2	1	4	3	4
	Total	2	2	7	8	11
CD001218	Czech			4	1	1
	German			2	2	1
	Polish			2		2
	Romanian		1			2
	Total		1	8	3	6
CD003566	Czech		3	1	2	
	German	1	1	1		
	Polish			2		2
	Romanian			1	1	3
	Total	1	4	5	3	5
CD003932	Czech		1	3	1	1
	German			2	1	2
	Polish			2		1
	Romanian					2
	Total		1	7	2	6
CD004344	Czech			1	4	
	German			1	1	3
	Polish			2		
	Romanian		1			1
	Total		1	4	5	4
CD006420	Czech			3	1	
	German		1	1	1	1
	Polish			2		
	Romanian					2
	Total		1	6	2	3
CD007871	Czech			3	1	
	German			2	2	
	Polish			2		
	Romanian	1		1	1	6
	Total	1		8	4	6
	Polish			1	4	6
	Romanian					1
	Total			3	14	13
CD011475	Czech			2	2	1
	German		1	6	5	1
	Polish			2	4	3
	Romanian			1		5
	Total		1	11	11	10
CD011693	Czech			3	2	
	German		2	2		4
	Polish		1		2	5
	Romanian					2
	Total		3	5	4	11
CD011714	Czech			2	1	1
	German			2	3	1
	Polish			1	1	3
	Romanian				1	1
	Total			5	6	6
CD011814	Czech				3	2
	German			1	4	3
	Polish		1	1	1	1
	Romanian			1		3
	Total		1	3	8	9
CD011826	Czech			1	1	3
	German			2	3	1
	Polish			1		1
	Romanian	1	1		4	9
	Total	1	1	4	8	14

CD008011	Czech			2	4	
	German		1		3	2
	Polish		1		1	1
	Romanian	1	1	3	1	6
	Total	1	3	5	9	9
CD009678	Czech				3	3
	German		1	1	3	1
	Polish		1	1		4
	Romanian	1		2	1	3
	Total	1	2	4	7	11
CD010227	Czech				5	2
	German			2	2	4
	Polish				1	3
	Romanian	1	2	2	5	8
	Total	1	2	4	13	17
CD010243	Czech			1	1	3
	German				4	3
	Polish			2		3
	Romanian					3
	Total			3	5	12
CD010531	Czech		1	3	1	
	German	1		2	1	2
	Polish		1	1		2
	Romanian					2
	Total	1	2	6	2	6
CD010726	Czech				3	1
	German		1	1	3	2
	Polish				2	4
	Romanian		1	2		1
	Total		2	3	8	8
CD010982	Czech			1	3	3
	German			3	8	5
	Polish			1	6	6
	Romanian	1	1		1	4
	Total	1	1	5	18	18
CD011319	Czech				4	
	German			2	6	6

Figure 6: Breakdown of Y2 machine translations survey results by PLS (CD number), language and star ratings. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy

CD number	Language	5 stars	4 stars	3 stars	2 stars	1 star
CD001088	German	1	1			
	Polish	6	2			
	Total	7	3	1		
CD001188	German		1			
	Polish	5	1			
	Total	5	2			
CD005050	German	1				
	Polish	3	2			
	Total	4	2			
CD005542	German	2	1			
	Polish	3	1	1		
	Total	5	2	1		
CD005576	German	3				
	Polish	2	1	1		
	Total	5	1	1		
CD006170	German		2	1		
	Polish	4	1			
	Total	4	3	1		
CD006941	German	2	2			
	Polish	2	2			
	Total	4	4			
CD007825	German	2		2		
	Polish	8		2		
	Total	10		4		
CD009002	German	9	2		1	
	Polish	6				
	Total	15	2		1	
CD010070	German		2			
	Polish	3	2			
	Total	3	5			
CD010182	German	1		1		
	Polish	3	2			
	Total	4	2	1		
CD010607	German		1			
	Polish	4		2		1
	Total	4	1	2		1
CD010697	German	2	1			
	Polish	5	1			
	Total	7	2			
CD010735	German	2	1			
	Polish	5				
	Total	7	1			
CD010743	German	3				
	Polish	4	1	1		
	Total	7	1	1		
CD011017	German	1		1		
	Polish	2	2			
	Total	3	2	1		
CD011045	German	2		1		
	Polish	3	1	1		
	Total	5	1	2		
CD011134	German	3	1	1		
	Polish	2	1	1		
	Total	5	2	2		
CD011694	German	8		1		
	Polish	4	1			
	Total	12	1	1		
CD011834	German	2				
	Polish	5	1			
	Total	7	1			

Figure 7: Breakdown of volunteer translation survey results by PLS (CD number), language and star ratings. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy.

CD number	Language	5 stars	4 stars	3 stars	2 stars	1 star
CD000501	Czech			1		
	German		4		1	
	Polish				2	
	Romanian			1	2	10
	Total		4	2	5	10
CD001218	Czech			1	1	
	German		4	1	2	
	Polish			2		
	Romanian			1		2
	Total		4	5	3	2
CD003566	Czech		1			
	German	1	2	1		
	Polish				1	
	Romanian			2	6	
	Total	1	3	3	7	
CD003932	Czech			1		
	German			1	2	
	Polish			1		
	Romanian			1	1	1
	Total			4	3	1
CD004344	Czech					
	German		1	1	1	1
	Polish		1	1		
	Romanian		1	1	2	2
	Total		3	3	3	3
CD006420	Czech					
	German		2	1	2	
	Polish		1	2		
	Romanian	1			3	3
	Total	1	3	3	5	3
CD007871	Czech	1				
	German		1	1	1	2
	Polish			1	1	
	Romanian			1	2	4
	Total	1	1	3	4	6
CD008011	Czech					
	German		4		1	
	Polish			1	1	
	Romanian	1			2	12
	Total	1	4	1	4	12
CD009678	Czech				2	
	German		1	2	1	
	Polish			1	1	
	Romanian				1	3
	Total		1	3	5	3
CD010227	Czech		1	1		
	German		1	3		
	Polish		1	1		
	Romanian		2	11	20	35
	Total		5	16	20	35
CD010243	Czech					
	German			2	1	
	Polish			2		
	Romanian				1	
	Total			4	2	
CD010531	Czech					
	German	1	3	1		
	Polish		2			1
	Romanian					4
	Total	1	5	1		5
CD010726	Czech		2			
	German		3	4		
	Polish			1	2	
	Romanian				1	2
	Total		5	5	3	2
CD010982	Czech			1	1	
	German		1	9	2	1
	Polish	1	2	1	1	2
	Romanian			1	3	6
	Total	1	3	12	7	9
CD011319	Czech					
	German		1	5	1	1
	Polish		1	1		2
	Romanian			1	1	1
	Total		2	7	2	4
CD011475	Czech					
	German		4	1		1
	Polish			1	1	1
	Romanian				2	1
	Total		4	2	3	3
CD011693	Czech					
	German		2	2	1	
	Polish			3		1
	Romanian				1	
	Total		2	5	2	1
CD011714	Czech					
	German		1	2	2	
	Polish					2
	Romanian				1	
	Total		1	2	3	2
CD011814	Czech					
	German			3	3	
	Polish			1	1	
	Romanian		1	2	4	7
	Total		1	6	8	7
CD011826	Czech		1			
	German		2	2		1
	Polish				1	1
	Romanian	1			9	30
	Total	1	3	2	10	32

Figure 8: Breakdown of Y2 neural machine translations survey results by PLS (CD number), language and star ratings. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy.

CD number	Language	Yes, more useful	No, not more useful	Question skipped
CD000501	Czech	1		
	German	3	1	1
	Polish		2	
	Romanian	3	9	1
	Total	7	12	2
CD001218	Czech	2		
	German	4	3	
	Polish		2	
	Romanian		3	
	Total	6	8	
CD003566	Czech	1		
	German	2	2	
	Polish		1	
	Romanian		4	4
	Total	3	7	4
CD003932	Czech	1		
	German	1	2	
	Polish		1	
	Romanian	1	2	
	Total	3	5	
CD004344	Czech			
	German	1	3	
	Polish		2	
	Romanian		6	
	Total	1	11	
CD006420	Czech			
	German	3	2	
	Polish		3	
	Romanian	2	3	2
	Total	5	8	2
CD007871	Czech	1		
	German	1	4	
	Polish		2	
	Romanian	1	5	1
	Total	3	11	1
CD008011	Czech			
	German	4	1	
	Polish		2	
	Romanian	4	8	3
	Total	8	11	3
CD009678	Czech	1	1	
	German	1	3	
	Polish		2	
	Romanian		4	
	Total	2	10	
CD010227	Czech	1	1	
	German	3	1	
	Polish		1	1
	Romanian	13	29	26
	Total	17	32	27
CD010243	Czech			
	German	2	1	
	Polish		2	
	Romanian		1	
	Total	2	4	
CD010531	Czech			
	German	3	2	
	Polish		2	1
	Romanian		4	
	Total	3	8	1
CD010726	Czech	1	1	
	German	5	2	
	Polish		3	
	Romanian	1	1	1
	Total	7	7	1
CD010982	Czech	2		
	German	8	4	1
	Polish	2	6	1
	Romanian		7	1
	Total	12	17	3
CD011319	Czech			
	German	4	4	
	Polish		4	
	Romanian		3	
	Total	4	11	
CD011475	Czech			
	German	5	1	
	Polish		3	
	Romanian		3	
	Total	5	7	
CD011693	Czech			
	German	3	2	
	Polish		4	
	Romanian		1	
	Total	3	7	
CD011714	Czech			
	German	4	1	
	Polish		2	
	Romanian		1	
	Total	4	4	
CD011814	Czech			
	German	2	3	1
	Polish		2	
	Romanian	3	11	
	Total	5	16	1
CD011826	Czech		1	
	German	2	3	
	Polish		2	
	Romanian	6	20	14
	Total	8	26	14

Figure 9: Breakdown of Y2 neural machine translations survey results: Responses by language whether the machine translation was more useful than only seeing the English text.

CD number	Language	5 stars	4 stars	3 stars	2 stars	1 star
CD001088	German	2				
	Polish	1				
	Total	3				
CD001188	German		1			
	Polish	1				
	Total	1	1			
CD005050	German		1			
	Polish	1				
	Total	1	1			
CD005542	German	3	2			
	Polish	1				
	Total	4	2			
CD005576	German		1			
	Polish		1			
	Total		2			
CD006170	German		1			
	Polish	1	1			
	Total	1	2			
CD006941	German	2				
	Polish		1			
	Total	2	1			
CD007825	German	2	2			
	Polish	1				
	Total	3	2			
CD009002	German	1	1			
	Polish	1	1			
	Total	2	2			
CD010070	German	2				
	Polish		1			
	Total	2	1			
CD010182	German	2		1	1	
	Polish	1				
	Total	3		1	1	
CD010607	German	2				
	Polish	1				
	Total	3				
CD010697	German		1			
	Polish	1	1			
	Total	1	2			
CD010735	German	1				
	Polish	1				
	Total	2				
CD010743	German	1	1			
	Polish	1				
	Total	2	1			
CD011017	German	1				
	Polish	1				
	Total	2				
CD011045	German		1			
	Polish	2				
	Total	2	1			
CD011134	German	4	1			
	Polish	1				
	Total	5	1			
CD011694	German	1				
	Polish	1				
	Total	2				
CD011834	German	1				
	Polish	1				
	Total	2				

Figure 10: Breakdown of volunteer translation survey results, second survey, by PLS (CD number), language and star ratings. 1 = very hard to understand, 2 = hard, 3 = neutral, 4 = easy, 5 = very easy.

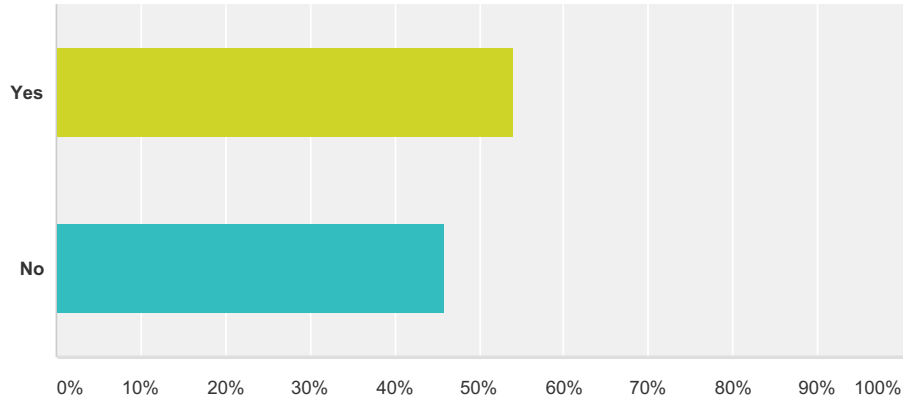
CD number	Language	Yes, more useful	No, not more useful	Question skipped
CD001088	German	1	1	
	Polish		1	
	Total	1	2	
CD001188	German		1	
	Polish		1	
	Total		2	
CD005050	German		1	
	Polish		1	
	Total		2	
CD005542	German	3	2	
	Polish		1	
	Total	3	3	
CD005576	German		1	
	Polish		1	
	Total		2	
CD006170	German		1	
	Polish	1		1
	Total	1	1	1
CD006941	German	1	1	
	Polish		1	
	Total	1	2	
CD007825	German	3	1	
	Polish		1	
	Total	3	2	
CD009002	German	1	1	
	Polish	1	1	
	Total	2	2	
CD010070	German	1	1	
	Polish		1	
	Total	1	2	
CD010182	German	3	1	
	Polish		1	
	Total	3	2	
CD010607	German	1	1	
	Polish		1	
	Total	1	2	
CD010697	German		1	
	Polish	1	1	
	Total	1	2	
CD010735	German		1	
	Polish		1	
	Total		2	
CD010743	German	1	1	
	Polish			1
	Total	1	1	1
CD011017	German		1	
	Polish		1	
	Total		2	
CD011045	German		1	
	Polish	1	1	
	Total	1	2	
CD011134	German	2	1	2
	Polish		1	
	Total	2	2	2
CD011694	German		1	
	Polish		1	
	Total		2	
CD011834	German		1	
	Polish		1	
	Total		2	

Figure 11: Breakdown of volunteer survey results, second survey: Responses by language whether the machine translation was more useful than only seeing the English text.

B Results NHS24 User Survey

Q1 Have you used the internet to access Scottish Health Information before?

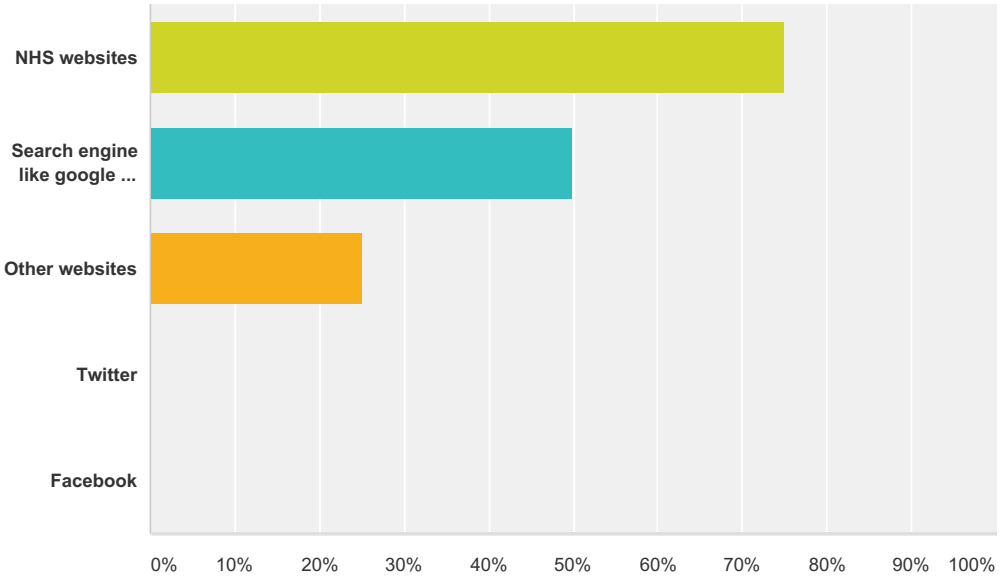
Answered: 24 Skipped: 0



Answer Choices	Responses	
Yes	**54.17%**	13
No	**45.83%**	11
Total		**24**

Q2 How did you access this information? Tick all that apply

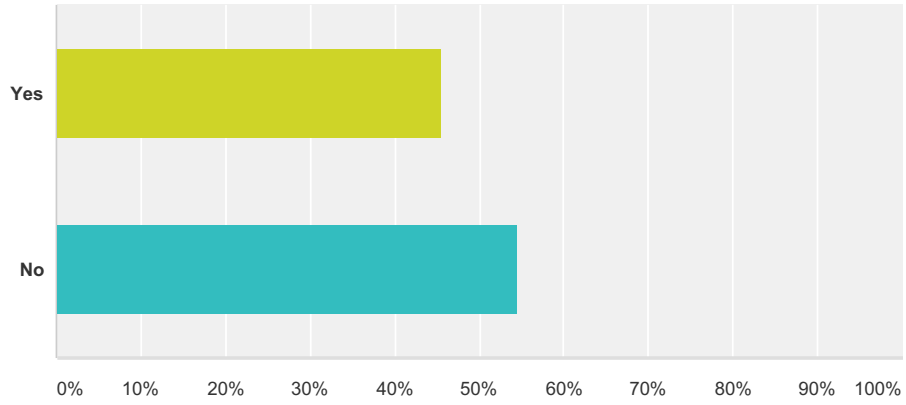
Answered: 8 Skipped: 16



Answer Choices	Responses	
NHS websites	**75.00%**	6
Search engine like google or bing	**50.00%**	4
Other websites	**25.00%**	2
Twitter	**0.00%**	0
Facebook	**0.00%**	0
Total Respondents: 8		

Q3 Have you had difficulty accessing health information in your own language when in Scotland?

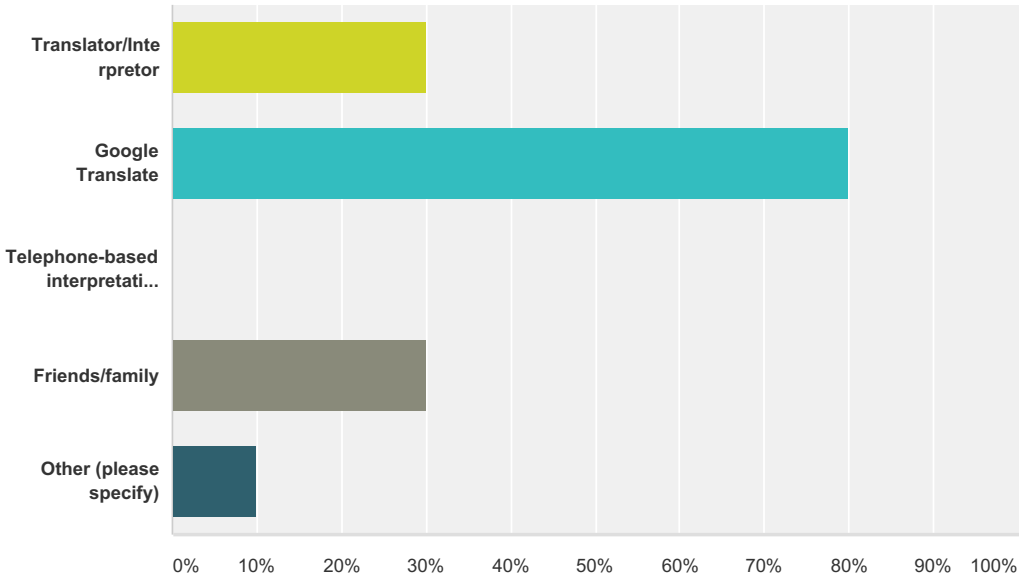
Answered: 11 Skipped: 13



Answer Choices	Responses	
Yes	**45.45%**	5
No	**54.55%**	6
Total		**11**

Q4 Have you used any of the following translation aids when using online or face to face services?

Answered: 10 Skipped: 14

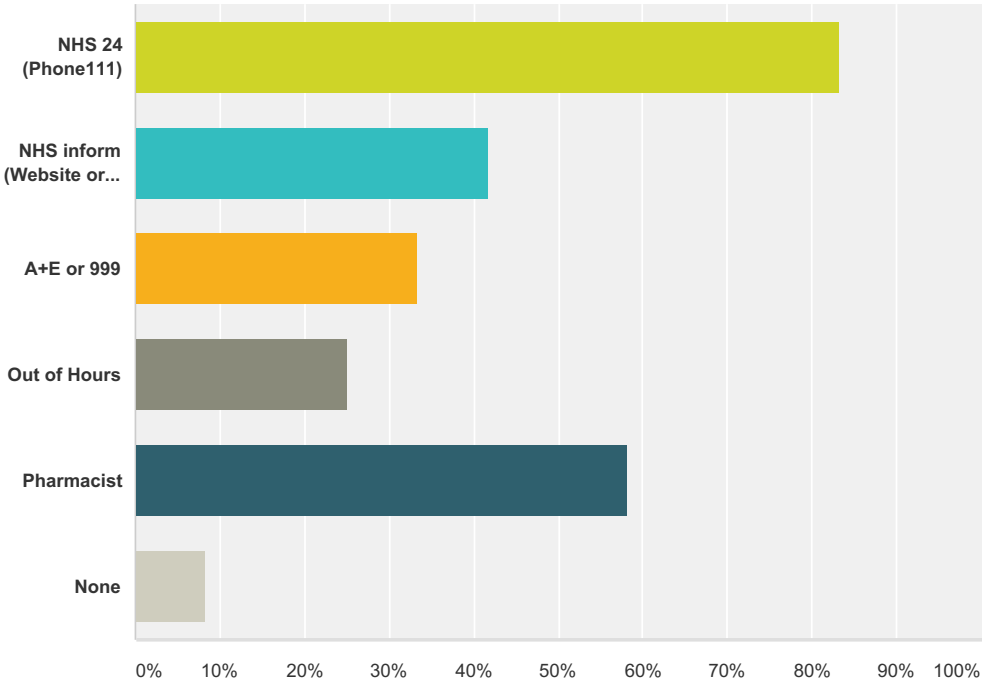


Answer Choices	Responses	
Translator/Interpretor	**30.00%**	3
Google Translate	**80.00%**	8
Telephone-based interpretation service	**0.00%**	0
Friends/family	**30.00%**	3
Other (please specify)	**10.00%**	1
Total Respondents: 10		

#	Other (please specify)	Date
1	None	4/9/2017 11:32 PM

Q5 Would you use any of the following services when your GP surgery is closed?

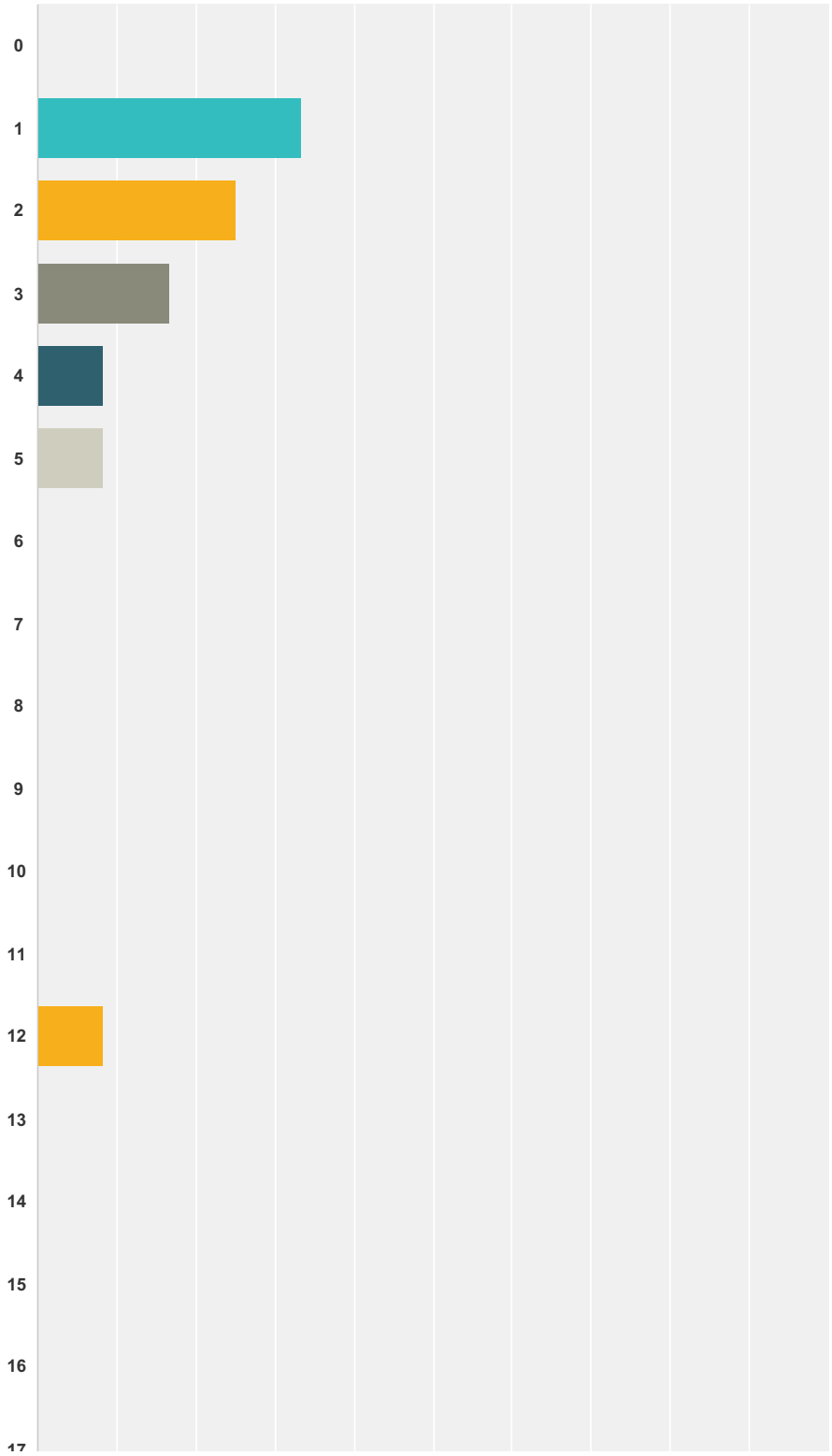
Answered: 12 Skipped: 12

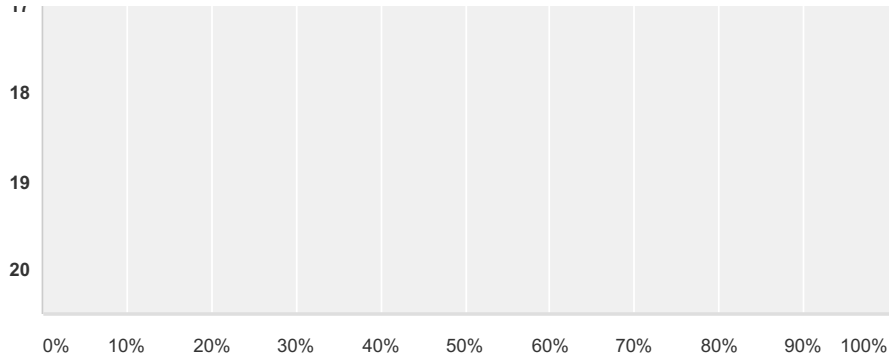


Answer Choices	Responses	
NHS 24 (Phone111)	**83.33%**	10
NHS inform (Website or phone)	**41.67%**	5
A+E or 999	**33.33%**	4
Out of Hours	**25.00%**	3
Pharmacist	**58.33%**	7
None	**8.33%**	1
Total Respondents: 12		

Q6 In the past 12 months how many times have you seen a doctor or nurse, been to a hospital or clinic, or used another NHS Scotland service?

Answered: 12 Skipped: 12

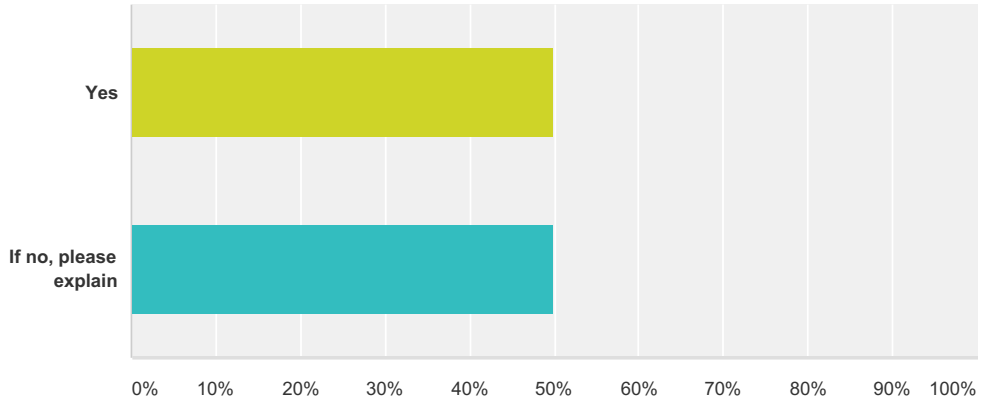




Answer Choices	Responses	
0	**0.00%**	0
1	**33.33%**	4
2	**25.00%**	3
3	**16.67%**	2
4	**8.33%**	1
5	**8.33%**	1
6	**0.00%**	0
7	**0.00%**	0
8	**0.00%**	0
9	**0.00%**	0
10	**0.00%**	0
11	**0.00%**	0
12	**8.33%**	1
13	**0.00%**	0
14	**0.00%**	0
15	**0.00%**	0
16	**0.00%**	0
17	**0.00%**	0
18	**0.00%**	0
19	**0.00%**	0
20	**0.00%**	0
Total		**12**

Q7 If you had access to health information in your own language would this have stopped you from visiting a service in person?

Answered: 8 Skipped: 16

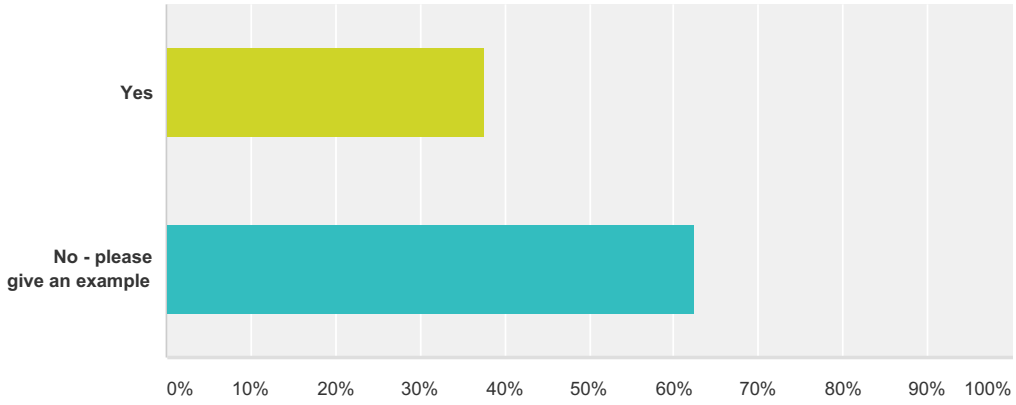


Answer Choices	Responses	
Yes	**50.00%**	4
If no, please explain	**50.00%**	4
Total		**8**

#	If no, please explain	Date
1	the answer to this question depends on the kind of information I'd access	4/6/2017 1:05 PM
2	If I was worried I would like to visit a doctor in person. Most of my appointments were related to my child's vaccinations etc. and they had to be attended in person. However I can imagine that if someone was not able to access NHS information online in English, they would be more likely to consult a helthcare provider.	4/5/2017 3:46 PM
3	I will feel insure	4/5/2017 12:03 PM
4	i needed meds	4/5/2017 10:45 AM

Q8 Is the translation accurate?

Answered: 8 Skipped: 16

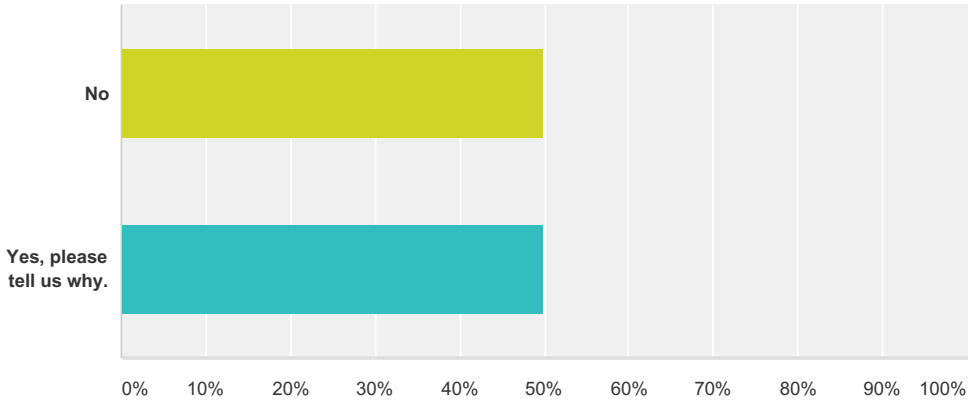


Answer Choices	Responses	
Yes	**37.50%**	3
No - please give an example	**62.50%**	5
Total		**8**

#	No - please give an example	Date
1	many sentences doesn't make sense, are confusing or written in the language nobody's using anymore	4/8/2017 5:36 PM
2	not all the words in the Polish translation are Polish, repetition of words, very literal translation	4/6/2017 1:05 PM
3	Sometimes there are a few incorrect words/phrases used in the wrong context which makes the website harder to understand.	4/5/2017 4:32 PM
4	pasożytować dziecka z zimną wodę - I think you meant to discourage sponging with cold water, but pasożytować means "being a parasite". There are a few similar examples, but in general the text is readable and most infomation is accurate.	4/5/2017 3:46 PM
5	grammatical	4/5/2017 12:03 PM

Q9 Are any of the words or phrases used wrongly or inappropriately in the context of health?

Answered: 8 Skipped: 16

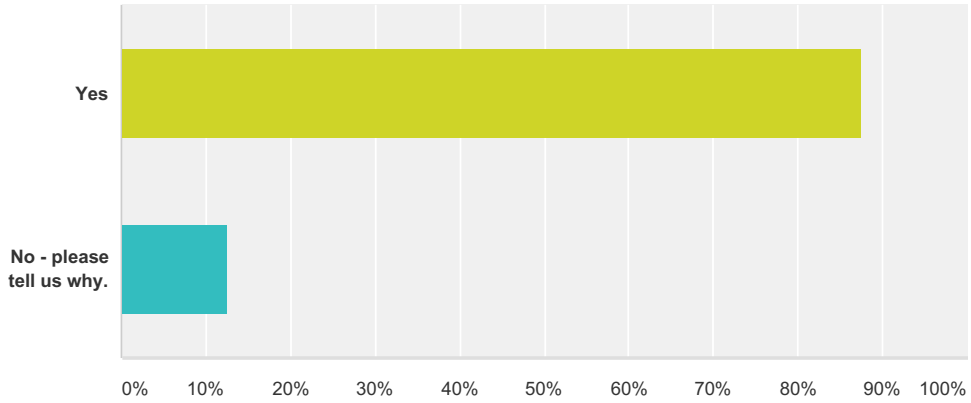


Answer Choices	Responses	
No	**50.00%**	4
Yes, please tell us why.	**50.00%**	4
Total		**8**

#	Yes, please tell us why.	Date
1	I cannot give you the exact example, but I remember many instances especially with instructions on how to correctly use equipment or medication or exercise, there's been multiple words used wrongly. This made it more difficult and confusing. On some occasions you also use words that are not used by Polish speakers on daily basis. Again, very confusing.	4/8/2017 5:36 PM
2	Spelling mistakes in the English version! (langauge??) The words cannot be translated into other languages if the original words are misspelled...	4/6/2017 1:05 PM
3	Incorrect spelling, incorrect punctuation, words and phrases used in the wrong context	4/5/2017 4:32 PM
4	1. Wirus B-limfotropowy - czyli wirusem wywołującym gorączkę i wysypkę - not sure what 'Wirus B-limfotropowy' is. Meningitis? 2. Wspólne dziecinnej choroby, takie jak ospę wietrzną i koklusz - Wspólne dziecinnej choroby? not sure what that means 3. odmowa paszy, floppiness lub senność. - floppiness not translated, "paszy or pasza" means food for animals 4. mleko matki lub wzoru - wzoru/wzór means mathematical equation, but here should read "mleko modyfikowane" i.e. formula milk 5. - zapalenie płuc, zapalenie tkanki płuc, zazwyczaj wywoływaną przez zakażenie - are there some words missing here?	4/5/2017 3:46 PM

Q10 Does the website contain information that may be useful to you and your friends and relatives in Scotland?

Answered: 8 Skipped: 16



Answer Choices	Responses	
Yes	**87.50%**	7
No - please tell us why.	**12.50%**	1
Total		**8**

#	No - please tell us why.	Date
1	the webite does not contain any information apart from the welcome section	4/6/2017 1:05 PM

References

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Birch, A., Abend, O., Bojar, O., and Haddow, B. (2016). HUME: Human UCCA-based evaluation of machine translation. *arXiv preprint arXiv:1607.00030*.

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2003). The Prague dependency treebank. In *Treebanks*, pages 103–127. Springer.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.

Bojar, O., Graham, Y., and Kamran, A. (2017). Results of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Tasks Papers*, Copehagen, Denmark. Association for Computational Linguistics.

Mareček, D., Bojar, O., Hübsch, O., Rosa, R., and Variš, D. (2017). CUNI experiments for WMT17 metrics task. To appear in Proceedings of WMT17.

Nivre, J., Agić, Ž., Ahrenberg, L., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., Bauer, J., Bengoetxea, K., Berzak, Y., Bhat, R. A., Bosco, C., Bouma, G., Bowman, S., Cebiroğlu Eryiğit, G., Celano, G. G. A., Çöltekin, Ç., Connor, M., de Marneffe, M.-C., Diaz de Ilarraza, A., Dobrovoljc, K., Dozat, T., Droganova, K., Erjavec, T., Farkas, R., Foster, J., Galbraith, D., Garza, S., Ginter, F., Goenaga, I., Gojenola, K., Gokirmak, M., Goldberg, Y., Gómez Guinovart, X., Gonzáles Saavedra, B., Grūzītis, N., Guillaume, B., Hajič, J., Haug, D., Hladká, B., Ion, R., Irimia, E., Johannsen, A., Kaşıkara, H., Kanayama, H., Kanerva, J., Katz, B., Kenney, J., Krek, S., Laippala, V., Lam, L., Lenci, A., Ljubešić, N., Lyashevskaya, O., Lynn, T., Makazhanov, A., Manning, C., Mărănduc, C., Mareček, D., Martínez Alonso, H., Mašek, J., Matsumoto, Y., McDonald, R., Missilä, A., Mititelu, V., Miyao, Y., Montemagni, S., Mori, K. S., Mori, S., Muischnek, K., Mustafina, N., Müürisep, K., Nikolaev, V., Nurmi, H., Osenova, P., Øvrelid, L., Pascual, E., Passarotti, M., Perez, C.-A., Petrov, S., Piitulainen, J., Plank, B., Popel, M., Pretkalniņa, L., Prokopidis, P., Puolakainen, T., Pyysalo, S., Ramasamy, L., Rituma, L., Rosa, R., Saleh, S., Saulīte, B., Schuster, S., Seeker, W., Seraji, M., Shakurova, L., Shen, M., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Simov, K., Smith, A., Spadine, C., Suhr, A., Sulubacak, U., Szántó, Z., Tanaka, T., Tsarfaty, R., Tyers, F., Uematsu, S., Uria, L., van Noord, G., Varga, V., Vincze, V., Wang, J. X., Washington, J. N., Žabokrtský, Z., Zeman, D., and Zhu, H. (2016a). Universal dependencies 1.3. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016b). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association.

Och, F. J. and Ney, H. (2000). A Comparison of Alignment Models for Statistical Machine Translation. In *Proceedings of the 17th conference on Computational linguistics*, pages 1086–1090. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Popel, M. and Žabokrtský, Z. (2010). TectoMT: Modular NLP framework. In Loftsson, H., Rögnvaldsson, E., and Helgadottir, S., editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

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

Straka, M., Hajič, J., and Straková (2016). UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Paris, France. European Language Resources Association (ELRA).