# HimL: Health in my Language

**Barry Haddow[1], Alexandra Birch[1], Ondřej Bojar[2], Fabienne Braune[5], Colin Davenport[3], Alex Fraser[5], Matthias Huck[5], Michal Kašpar[6], Květoslava Kovaříková[6], Josef Plch[6], Anita Ramm[5], Juliane Ried[4], James Sheary[3], Aleš Tamchyna[2], Dušan Variš[2], Marion Weller[5], Phil Williams[1]**

[1]School of Informatics, University of Edinburgh, Scotland

[2] Faculty of Mathematics and Physics,Charles University in Prague, Czech Republic

[3] NHS 24, Caledonia House, Glasgow, Scotland [4] Cochrane, Freiburg, Germany

[5] LMU Munich, Germany [6] Lingea s.r.o., Brno, Czech Republic

Coordinator email: bhaddow@inf.ed.ac.uk

## Abstract

HimL (www.himl.eu) is a three-year EU H2020 Innovation Action, which started in February 2015. Its aim is to increase the availability of public health information via automatic translation. Targeting languages of Central and Eastern Europe (Czech, German, Polish and Romanian) we aim to produce translations which are adapted to the health domain, semantically accurate and morphologically correct.

## 1 Description

In HimL we aim to deploy and evaluate machine translation systems for the public health domain, addressing domain adaptation, semantic accuracy and target morphology. The project is now in its third year, and we have made two releases of our translation systems and used them to translate the user partner websites. These have been subjected to automatic evaluation, human evaluation, and are undergoing user evaluation.

The HimL system releases so far were built as phrase-based MT systems using large, diverse training sets and applying language model and translation model interpolation to adapt to the medical domain. In Year 2, we applied the corrective approach to morphology to the English→Czech system, and the two-step approach to the English→German system. We also filtered the phrase tables to remove phrase-pairs that would clearly result in semantically incorrect translations.

Our work on human evaluation has led us to develop a semantic evaluation measure based on the UCCA (Universal Conceptual Cognitive Annotation) framework. We are currently developing an automatic version of this metric to give rapid feedback on the semantic accuracy of translations.

We have recently shown that neural MT can produce better results for most of our language pairs, using continued training with synthetic data for adaptation, and will be rolling out NMT systems in Year 3. We are investigating how our work on semantic accuracy and treatment of morphology can be applied to NMT, for instance by incorporating semantic roles into the NMT system, or by using additional signal from back-translation to confirm the semantic accuracy. Our machine-learning version of the corrective morphology tool depfix (known as MLfix) will be used in the Year 3 system releases.

Finally, we are sponsoring this year's WMT biomedical translation task[1], providing test sets for the HimL language pairs, and collaborating in the release of a medical MT training set (UFAL Medical Corpus).

## Acknowledgement

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



[1]http://www.statmt.org/wmt17/biomedical-translation-task.html