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

D5.1: Test Set

Author(s): Alexandra Birch

Dissemination Level: Public

Due Date: October 1st 2015

Version 1.0

Grant agreement no.	644402
Project acronym	HimL
Project full title	HimL will deliver high quality machine translation for the public health domain that is semantically accurate and supports morphologically rich languages of Central and Eastern Europe.
Funding Scheme	Innovation Action
Coordinator	Barry Haddow (UEDIN)
Start date, duration	1 February 2015, 36 months
Distribution	Public
Contractual date of delivery	October 1,st 2015
Actual date of delivery	October 1,st 2015
Deliverable number	D5.1
Deliverable title	Test Set
Type	Other
Status and version	1.0
Number of pages	5
Contributing partners	UEDIN
WP leader	UEDIN
Task leader	UEDIN
Authors	Alexandra Birch
EC project officer	Martina Eydner
The Partners in HimL are:	The University of Edinburgh (UEDIN), United Kingdom
	Univerzita Karlova V Praze (CUNI), Czech Republic
	Ludwig-Maximilians-Universitaet Muenchen (LMU-MUNICH), 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 bhaddow@staffmail.ed.ac.uk
University of Edinburgh Phone: +44 (0) 131 651 3173

Contents

1 Introduction

This document provides a brief description of the HimLtest set which was created with data from NHS 24's website and from Cochrane reviews. This resource will be used throughout the project to validate our research efforts, and to tune the systems for optimal performance. This test set will soon be made freely available for research purposes.

2 General description of the Test Set

The HimL test set consists of in-domain data for tuning and testing from the use case partners, NHS 24 and Cochrane. The size of these data sets is about 30,000 English words for each use case partner, and the data is split evenly between tuning and testing sets. The NHS 24 data was scraped from a list of URLs and the Cochrane data was provided in XML formal. The texts were sentence segmented, some normalisation of punctuation was performed, duplicate sentences were eliminated, and some boilerplate text was removed. The clean English data was then translated into Czech, Polish, Romanian and German using professional translators. These translators used the source sentences and post-edited machine translations of the source, in a post-editing tool provided by Lingea. This was done to provide post-editing data to train automatic post-editing tools for other tasks in the project. The HimL data has been split between tuning and testing. We have careful to keep entire pages together, and to keep similar quantities of text across the two data sets. These data sets are available to the consortium via the version control system that is already in place. In Table 1 we can see the number of lines and the number of words that each part of the test set contains.

	Number of Lines	EN	DE	RO	PL	CS
NHS 24	2459	30198	30306	31622	28249	27834
Cochrane	1433	29944	30059	34416	32044	28680
Total	3892	60142	60355	66038	60283	56514

Table 1: Number of words for the HimL tuning and testing data

3 NHS 24

For NHS 24 we crawled a set of URLS from the NHS inform website for content which is in the style that NHS 24 considers to be the most representative of the kind of text across their entire website. Half of the HimL pages come from the "Falls" section of the website, and the other half come from a variety of conditions from the health library. In Table 2 we can see an example of the kinds of sentences contained in the NHS 24 section of the HimL test set.

Healthy diet - Falls prevention | NHS inform
How to eat well
Eating regular, nutritious meals and drinking plenty of non-alcoholic fluids can help to avoid problems that can contribute to a fall, including:
lightheadedness
depression.

Table 2: Example sentences from "Falls" section of the NHSInform web site

The text on NHS 24 website is written very carefully to be easy to read for the general public. Health literacy is a big concern in NHS Scotland, and people with low health literacy levels have been shown to experience much poorer outcomes. The sentences in this part of the test set are therefore mostly short and the vocabulary is as simple as possible, including the most commonly used forms for technical words relating to conditions, treatments, and anatomy. There is also a relatively large number of lists and short section titles, as the texts are quite structured to highlight important information.

4 Cochrane

Cochrane data is selected from reviews to cover a representative sample of the kinds of data that needs to be machine translated. This data is highly structured XML text that consists of two parts, a plain language summary and a technical abstract. In Table 3

we can see examples of kind sentences contained in the Cochrane section of the HimL test set.

About one in five people treated for breast cancer develop lymphoedema later on. We reviewed the available evidence to determine whether some methods, such as manual lymph drainage (a massage therapy), compression, exercise or only education could help prevent lymphoedema.

Table 3: Example sentences from a Cochrane plain language summary

These articles are aimed at medical professionals and so even the plain language summary text is of a much greater complexity than the NHS 24 text. The sentences are longer, and there are many technical terms and they contain specialised vocabulary. The abstracts are slightly more complex than the plain language summaries, but we have yet to determine if they constitute a genuinely different domain or not.

5 Conclusion

This deliverable has provided a short description of the HimL test set, which will be used to tune our systems to the health domain, and to validate our research agenda.