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

D5.2: First Evaluation Report

Author(s): Alexandra Birch, Barry Haddow, Ondřej Bojar, Juliane Ried, Colin Davenport

Dissemination Level: Public

Date: February, 1st 2016

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	February, 1st 2016
Actual date of delivery	February 1st 2016
Deliverable number	D5.2
Deliverable title	First Evaluation Report
Type	Report
Status and version	1.0
Number of pages	17
Contributing partners	UEDIN
WP leader	UEDIN
Task leader	UEDIN
Authors	Alexandra Birch, Barry Haddow, Ondřej Bojar, Juliane Ried, Colin Davenport
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-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 bhaddow@staffmail.ed.ac.uk
University of Edinburgh Phone: +44 (0) 131 651 3173

Contents

1 Executive Summary **4**

2 HimL Test Set and Evaluation **4**

2.1 HimL Test Set . . . 4
2.2 HimL Systems . . . 4
2.3 Applying In-Domain Data to HimL Systems . . . 5

3 Evaluation Plan **6**

3.1 Overview . . . 6
3.2 Fine-Grained Semantic Analysis . . . 7
3.3 Ranking . . . 7
3.4 Gap-filling . . . 7
3.5 Post-Editing Speedup . . . 8
3.6 User Acceptance Testing . . . 8
3.7 Evaluation of Impact . . . 8

4 Human Semantic Evaluation **9**

4.1 Overview . . . 9
4.2 Semantic Annotation . . . 9
4.3 MT Evaluation Overview . . . 10
4.3.1 Lexical nodes . . . 11
4.3.2 Structural nodes . . . 11
4.4 Machine Translation Evaluation Tool . . . 13
4.5 Results . . . 14
4.5.1 Overall Statistics . . . 14
4.5.2 IAA . . . 15
4.5.3 UCCA Scores . . . 16
4.6 Discussion . . . 17
4.7 Automatic Semantic Evaluation . . . 17

5 Outlook **17**

References **17**

1 Executive Summary

This document provides a summary of the work done in the first year of the HimL project in the evaluation work package 5. The objectives of this work package are to develop human and automatic accuracy-based evaluation strategies for machine translation, and to carefully evaluate the quality and impact of the innovations we deliver to NHS24 and Cochrane.

The first section in this report describes the creation of the HimL test set from our use case partners. We then describe experiments where we apply these test sets to the HimL machine translation systems, and we report the first results for our models on in-domain data.

The next section describes the evaluation plan which we developed in close collaboration with NHS24 and Cochrane. Work package 5 is responsible for the thorough evaluation of translation systems. Although automatic metrics give indications of the quality of the translations, they are not good at capturing accuracy or usefulness. Accuracy needs to be investigated using human evaluation which is very time consuming and expensive. The evaluation plan allows us to map out the schedule of these evaluations and defines the types of evaluation that both the academic partners and the use case partners will most benefit from.

The last section of this report describes the development of our human semantic evaluation method. This evaluation uses semantic trees on the source sentence and asks bilingual annotators to label which parts of the tree have been correctly translated. In this way we are able to quantify how much of the meaning of the source sentence has been retained in the translation. Having multiple annotators for two language pairs allowed us to investigate inter-annotator agreement, and results show that this approach is reliable. This human evaluation comprises the first year's user acceptance testing.

Task	Description	Planned Schedule	Status
5.1	Test corpora for the required language pairs	M1-M6	Complete
5.2	Human Semantic MT Evaluation Metric	M1-M12	Complete
5.3	Automatic Semantic MT Evaluation Metric	M6-M18	Initiated
5.4	User acceptance testing	M9-M12 annually	Year 1 testing completed as part of Task 5.2
5.5	Evaluation of the impact	M6-M36	Planning complete, started data collection

Table 1: Tasks in workpackage 5, their dates and their status as of 31/Jan/2016

Table 1 summarises the status of the tasks in the evaluation workpackage. We can see that all tasks are currently on schedule and that puts us in a strong position to succeed in delivering the work promised in year two and three of the HimL project.

2 HimL Test Set and Evaluation

The first priority of the project was to create a HimL test set which now constitutes an essential building block for our translation experiments in the consumer health domain. The creation of the test set was fully documented in the deliverable D5.1 which was submitted at the end of September 2015. The first year translation systems for HimL were also released at the end of September and they are described in the deliverable D4.1. In this section we will provide a brief summary of this work and we document how the HimL systems performed when tested on in-domain data.

2.1 HimL 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 (supplied by NHS 24) and the Cochrane data was provided in XML format. 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 help with other tasks in HimL, in particular the development of MLfix in task T3.3. The HimL data has been split between tuning and testing. We have been careful to keep entire pages together in the same part of the split, 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.

2.2 HimL Systems

In this section we describe the translation models which we created for the year 1 HimL system release. These systems consisted of phrase-based Moses systems, built using a similar approach to UEDIN's recent WMT (Workshop in Statistical Machine

Translation) systems. We trained these systems on all the freely available parallel training data that we could find for system building, which included the data released for the WMT15 news translation task, the Khresmoi-sponsored WMT14 medical translation task, and the data available from OPUS4. When tuning these models, the HimL test set was not yet available and we had to use alternatives from the consumer health domain. The En→Ro tuning set consisted 3000 sentences from EMEA, and for the other language pairs we used a combination of the tuning set of the WMT14 medical task and Cochrane translation memories for En→De and En→Pl.

The set of models and additional features that we employed were similar to those developed over the past few years in UEDIN in the context of the WMT evaluation campaigns. In particular, the models used were:

- 5-gram Kneser-Ney-smoothed language model, trained on each corpus individually, then linearly interpolated to minimise perplexity on the development set
- Phrase translation model, trained on concatenation of all parallel data, smoothed using Good-Turing. This contributes the standard 4 features, i.e forward and reverse phrase-pair probabilities, plus forward and reverse lexical scores.
- Hierarchical reordering model, predicting monotone, swapped, left and right reorderings.
- 5-gram operation sequence model (OSM)

For tuning we use k-best MIRA on the development set, selecting the weights that give best results on that set. In decoding we use a large cube-pruning pop limit (5000), as well as minimum Bayes risk decoding, and the "no reordering over punctuation" constraint.

2.3 Applying In-Domain Data to HimL Systems

The first system release of HimL occurred simultaneously to the creation of the test set, and therefore we were unable to report results on NHS24 and Cochrane data in deliverables D4.1 or in D5.1. In this section we report the first results of HimL systems tested on in-domain data from our use case partners.

System	HimL Cochrane	HimL NHS24
En→De	36.46	30.02
En→Cs	24.58	19.93
En→Ro	34.24	30.81
En→Pl	16.03	23.06

Table 2: Translation quality on HimL Test sets, measured in case-sensitive BLEU

In Table 2 we can see BLEU score results for the systems when tested with Cochrane abstracts and text from the NHS24 website. For most of the languages, we see better results for the Cochrane data than for the NHS24. En→Pl is the only language pair which shows higher BLEU scores for NHS24. This is a surprising result as NHS24 sentences tend to be short and the language is aimed at the general public which includes many people with low reading levels. The reason that results are higher for Cochrane could be that there are many lists and imperatives in the NHS24 test set, and these would not occur in the training data. In year 2 of the project we intend to thoroughly investigate the applicability of our systems to the consumer health domain, and adapt them accordingly.

3 Evaluation Plan

This section will summarise the HimL evaluation plan which focuses on the user acceptance part of the work plan. In order for us to apply costly human evaluations in the best possible manner, we need to plan the overall human evaluation strategy very carefully. The main focus of evaluation in HimL is to determine how well the machine translation systems perform in the context of the use cases. For this reason NHS24 and Cochrane are heavily involved in designing and participating in these tests. The evaluation plan is a work in progress, and will be updated regularly with contributions from other partners, and as plans for evaluation become more defined.

3.1 Overview

As HimL is an innovation action, we need to conclusively demonstrate that our models deliver appropriate high quality technology to NHS24 and Cochrane, and to users of their information services. We plan to do this using a small set of human evaluations which will aim to answer the following questions:

- Have we improved the accuracy of translation models?
- Do people prefer our translations to baseline system translations?
- How much useful information do the automatic translations supply to the end users?
- Are end users able to find and navigate the translations?
- Are end users' expectations of the translations well managed?
- In cases where automatic translations are not acceptable, does post-editing machine translation speed up human translation?

The annotators for human translation are recruited by the use case partners. They are native speakers of the HimL target languages ie. Polish, Romanian, Czech and German and in most cases they have a good level of English understanding as well. For some tasks annotators who have little or no English might be more appropriate as these are our target users.

Here is a summary of the individual human evaluations which we plan to perform:

- **Fine-grained semantic analysis:** The source English sentences are annotated with their essential semantic components, and then the accuracy of the aligned components in the machine translation will be evaluated. This analysis is slow and labour intensive but it results in gold standard data which can be used to evaluate automatic semantic metrics. This type of analysis refers to the human evaluation described in Section 4.
- **Ranking:** Annotators will be shown between two and five machine translations and they will also be shown the gold standard human translation. They will then be asked to rank machine translations according to their quality.
- **Gap-filling:** Annotators are shown a machine translated paragraph and then they will be shown a human translated summary sentence where one or more meaning bearing words are redacted. They are then asked to fill in the gap. Their accuracy in filling the missing words will indicate how well they understood the machine translated text.
- **Post-Editing Speedup:** Cochrane translators will post-edit machine translations. Their translation speed will be compared to when they post-edit baseline machine translations and to full manual translations.
- **User Acceptance Testing:** Usability surveys of NHS24 and Cochrane's websites will be conducted.
- **Impact Assessment:** Website traffic statistics will be continuously collected on web pages with automatic machine translations.

In Table 3 we can see the overall plan for human evaluation during the three years of the project. It is estimated that the amount of annotation for each human evaluation task, for each target language will be approximately a week's worth of effort, or around 40 hours. This may vary depending on availability of annotators and on the final experimental design. Although in principle NHS 24 is primarily responsible for Polish and Romanian evaluations, and Cochrane for German and Czech, for some evaluations user partners would provide annotators for all four language pairs in order to complete evaluations in a consistent and timely fashion.

Table 3: Evaluation effort for the HimL innovations and prototypes, broken down per year and per user partner

		Accuracy Eval.		Task Based		Field Testing	
		Fine	Rank	Gap	PE	User	Imp
Year 1	NHS 24	•	–	–	–	–	–
	Cochrane	•	–	–	–	–	–
Year 2	NHS 24	•	•	•	–	•	•
	Cochrane	•	•	•	•	•	•
Year 3	NHS 24	•	•	•	–	•	•
	Cochrane	•	•	•	•	•	•

Fine: Fine-grained Semantic Eval.; **Rank:** Ranking Eval. **Gap:** Gap Filling Eval.;
PE: PostEditing Speedup; **User:** User Acceptance Testing; **Imp:** Impact Assessment

3.2 Fine-Grained Semantic Analysis

We have been developing a human metric of adequacy, with the goal of creating gold standard evaluations which can be used to calibrate and to test automatic metrics. We base our approach on the human metric, HMEANT (Lo and Wu, 2011), by breaking down sentences into semantic units, thereby making evaluation more reliable. However, we extend HMEANT to include linkages between semantic units and remove its reliance on syntactic features. This will make evaluation more robust to natural variation between languages, and more complete. The annotation process consists of annotating the original English sentence with semantic trees. We will be using UCCA annoation for this (Abend and Rappoport, 2013) as UCCA has been developed based on typology studies over a large number of diverse languages to capture as far as possible universal semantic constants. English sentences with UCCA trees can be used for testing multiple translation systems. For a particular translation system, we will collect model output and have native-speaking human annotators step through each semantic component of the UCCA tree, to see if the aligned MT translation is correct or not. This way we will extract a fine-grained analysis of the number and type of the semantic components of the sentence which have been correctly translated.

3.3 Ranking

In the annual machine translation shared tasks in the Workshop on Machine Translation, the final quality results are provided by a human ranking task. Annotators are shown five machine translations from different translation systems and the gold standard human translation. They are then asked to rank machine translations according to their quality. This evaluation is fast and returns a large number of ranked pairs. We will perform this task to compare different versions of our machine translation models, and as a method for evaluating our human semantic metric.

3.4 Gap-filling

In this section we describe an evaluation which aims to detect how well annotators are able to perform a gisting task by getting them to fill in semantically important gaps in human translated sentences. These experiments will be similar to those described by Ageeva et al. (2015). The basic evaluation method is to ask informants to read MT output and then to fill gaps in the reference (human) translations. A designated number of keywords are removed from the human-translated sentences. The evaluators are then asked to fill the gaps with suitable words with the help of MT output. The gap-filling task models how well users comprehend the key points of the text, as it is roughly equivalent with answering questions. Thus, the method does not directly evaluate the quality of machine-produced text, but rather its usefulness in understanding the meaning of the original text.

In the context of Cochrane reviews, we might apply this evaluation strategy to human translated content defined by the PICO principle [1]

- P: Population or participants. Who are the relevant patients?
- I: The intervention or indicator. What is the strategy, test or exposure you are interested in?
- C: The comparator or the control. What is the control or alternative strategy, test or exposure?
- O: The outcome. What are the patient-relevant consequences of the intervention?

By testing the readers ability to complete PICO information, we will be able to gauge their grasp of the essential nature of the review summaries.

[1] http://learntech.physiol.ox.ac.uk/cochrane_tutorial/cochlibd0e84.php

In the context of NHS24, making sure that patients have understood the information that has been provided to them is essential to their organisation. Health literacy is a real challenge. There are tools and exercises that health professionals adopt in order to make sure that patients have understood their diagnosis and care instructions. The "chunk-and-check" method of communication is often used to determine whether patients have heard them correctly. After you have communicated one important message—a "chunk" of instructions—you check how much the patient understood, for example by asking "What will you tell your husband about how many physical therapy visits should be scheduled?". Together with NHS24, we could develop a series of "check" questions for accompanying different content articles.

3.5 Post-Editing Speedup

COCHRANE has extensive experience in translating content and they have always maintained a very high standard of accuracy. They use bilingual domain experts, i.e., clinicians, to translate their summary data as they have found that even professional translators make significant errors while translating medical health care documents. COCHRANE will therefore need to carefully evaluate the output of the HimL translation systems, to gauge whether or not they are of an acceptable standard to publish. This process will include allowing domain experts to post-edit MT output, potentially using the EU FP7 project MateCat 30 tools. Essentially, post-editing effort will be evaluated and the reduction of post-editing effort will be monitored. By the end of the project, our aim is that little or no post-editing would be required.

3.6 User Acceptance Testing

The idea of this task is to verify that the new translation functionality is properly managed and integrated into NHS24 and COCHRANE websites, before submitting it to the live websites. The evaluation will be performed by selected groups of users, and will assess whether the translations are of a high enough standard to be useful, whether the web site functionality is easy to navigate, and whether user's expectations about automatic machine translation are correctly managed. Since we are providing fully automatic translation there will still be errors and imperfections in translations, but this will not necessarily cause the translations to be rejected. Instead, we will carefully manage the users' expectations of the translations, installing suitable explanations of the technology and the rationale behind it.

NHS24 has had a policy of evaluating their multi-lingual content by use of community involvement. They contacted community groups which speak the languages which are most relevant to NHS24 Polish and Mandarin. Specifically, they identified a Polish online news site, GazetaE, and a Chinese community development project, CCDP, which were prepared to be involved in the evaluation of human translations. Reviewers were asked to rigorously evaluate translated content, and also to review the website's functionality. Examples of questions asked are:

- Is the translation accurate? Are any of the words or phrases used wrong or inappropriate in the health context?
- Is it easy to change from English into your language?
- Is it clear how to find the information you are looking for?
- Does the website contain information that is useful to you and your friends and relatives in Scotland?

In HimL, we will develop a survey for NHS24 website which reflects the fact that multilingual content is automatically and not manually translated. We will recruit Polish and Romanian speakers to evaluate the functionality, quality, and transparency of the translated content. COCHRANE will report their findings on the evaluation of the implementation of automatic translations on their website, and in particular, how well user expectations of automatic translations are managed.

3.7 Evaluation of Impact

We have defined the set of web-site statistics which NHS 24 and Cochrane will collect in order to track the impact of machine translation. As MT is not yet live on their main sites, the statistics are collected for comparative purposes to demonstrate the difference in traffic before and after applying MT.

NHS24 and Cochrane both use Google Analytics to track access to their respective websites. They have agreed to collect the following access statistics from January 2016 onwards on a monthly basis:

1. List of visitors by country, highlighting the rank of the countries of the HimL languages:
    - Number of sessions, new users, pages / session, average time on site, and bounce rate
2. Visits by browser languages (Google Analytics category Geo > Language):

- The same parameters as above, for each of the specific language codes: cz, pl, de, ro;
- Compared to the total sessions;
- Compared to the following languages: en, es, fr, it, pt, ja, nl, zh, ru, ar

3. Access to translated content in the four HimL languages (once MT content has been published):
   - The same parameters as above, for the full subdomain content of each HimL language respectively

Data from January to September 2016 will serve as a baseline. Following the publication of HimL MT year 2 content in September 2016, the expectation is that the access numbers would increase for the HimL languages and countries, and in comparison to other common languages, which would demonstrate the positive effect of the provision of native language content.

4 Human Semantic Evaluation

In HimL we focus on producing high accuracy machine translation systems, but common automatic MT metrics are not able to directly capture accuracy. Even previously suggested methods for using humans to evaluate accuracy are highly problematic. We aim to develop a human evaluation method which is reliable and affordable and apply it to the HimL MT prototypes. The work described in this section relates to task *T5.2: Human semantic evaluation*.

In November 2015, we ran an evaluation task with 6 bilingual annotators, 2 from NHS 24 and 4 from Cochrane. We asked them to annotate about 350 sentences translated with the HimL year one systems and they had a budget of up to 40 hours each to perform this task. In this section we motivate and describe the experiment that we ran and we provide an initial analysis of the results. In Year 2 and Year 3 we will refine this evaluation task and use it to track the progress of our HimL prototypes.

4.1 Overview

Semantic evaluation of machine translation has typically been done at the sentence level and we propose an approach which breaks down the evaluation into basic semantic units, making evaluation simpler and more consistent. Our assumption is that the semantic structure on the source sentence should be retained in the translation, and if it is not, then some essential part of the meaning is lost. The semantic framework that we base our evaluations on is called Universal Conceptual Cognitive Annotation (Abend and Rappoport, 2013). UCCA has been developed using linguistic theories about what types of components and structures are universal across many different languages.

Our goal is to quantify how much of the meaning of the source sentence is preserved through translation. There have been many approaches to evaluating the quality of machine translation, but most of them have asked the annotator to give a score for the entire sentence. There are of course many ways that a translation can be incorrect and asking an annotator to provide a global score for a sentence is a cognitively difficult task even if e.g. limited to a relative comparison with another candidate translation. How serious is an error? What is the impact of multiple errors on global meaning? By using UCCA structure to break the evaluation into meaningful components, we provide a more consistent and reliable method of evaluating translation accuracy.

The annotation proceeds as follows. Firstly, the source sentences (English, in our case) are annotated with UCCA trees. This annotation is normally performed by computational linguists, and requires some training in UCCA, but the annotation can be reused for different target languages and different MT systems. We then create translations of the source sentences with the MT system, collecting the word alignments from the source sentence to the translation provided by the system. These word alignments are used to project the UCCA annotation from the source sentence to the the translation output, and then bilingual annotators go through each projected UCCA node, assessing how well it is translated. We can estimate the impact of individual errors given their location in the semantic structure and we can thus extract a score for the whole sentence. More details on the procedure are provided below.

4.2 Semantic Annotation

The source sentences in this annotation scheme have been annotated with a semantic structure defined as Universal Conceptual Cognitive Annotation (UCCA). UCCA was developed in the Computational Linguistics Lab of the Computer Science Department of the Hebrew University by Omri Abend and Ari Rappoport. UCCA views the text as a collection of scenes (or events) and their inter-relations and participants.

As can be seen in Figure 1, the UCCA annotation results in a tree structure where each leaf is linked to a word in the sentence at the bottom. A scene must contain a process (P) or a state (S). It can also contain participants (A) and it can be linked to other scenes by a linker (Linker). Participants, processes and states can be further analysed into elaborators (E), centres (C) and relators (R). These labels are very high-level and relate to cognitive concepts which should remain stable across languages.

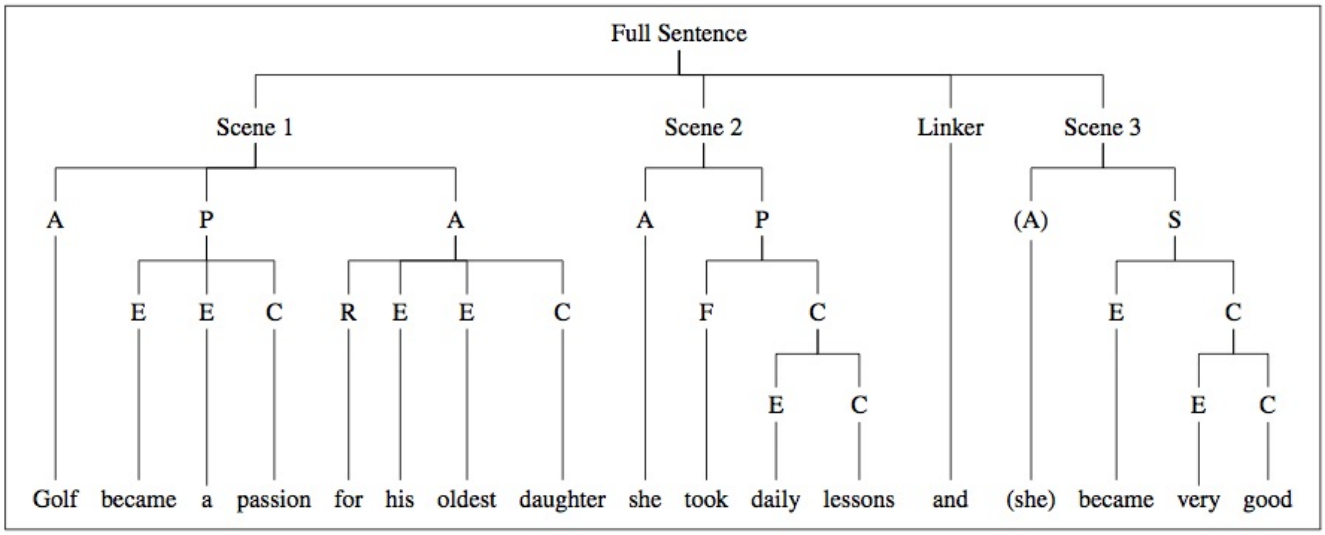

Figure 1: UCCA Tree with scenes

The fact that in UCCA the labels are cognitive concepts and that they are linked directly to words are both advantages when considering which semantic formalism is appropriate for machine translation evaluation.

One of such alternative formalisms is Abstract Meaning Representation Banarescu et al. (2013). AMR is being actively developed with a view towards using it as a way of generating translations but AMR graphs are not aligned to the words in the sentences. Having more abstract semantic structures makes the link between source words, target words, and structures more complex and potentially less useful. Furthermore, AMR has been developed mainly with English in mind, and it remains to be seen how universal AMR graphs are. See Xue et al. (2014) for first observations of divergences between English vs. Chinese and Czech AMRs.

Another possible semantic framework for this kind of MT evaluation is Semantic Role Labelling Palmer et al. (2010). SRL has been used in a human translation metric called HMEANT (Lo and Wu, 2011). HMEANT uses semantic role labels to measure how much of the "who, why, when, where" has been preserved in translation. Annotators are instructed to identify verbs as heads of semantic frames. Then they attach role fillers to the heads and finally they align heads and role fillers in the candidate translation with those in a reference translation. Using SRL for evaluating SMT has a number of disadvantages as explored by Birch et al. (2013) for German, Bojar and Wu (2012) for Czech and by Chuchunkov et al. (2014) for Russian. The most important drawbacks are as follows:

- SRL frames are based around a verb which is particularly problematic for copular verbs and when verbs are translated correctly as nouns or correctly omitted (the verb "to be" in some Russian constructions).
- SRL frames do not cover the entire source sentence and the semantic structure is therefore not completely defined, importantly links between frames are not considered and prepositional phrases which attach to nouns are not marked.
- Even considering a limited set of eleven roles (agent, patient, experiencer, locative etc.) is problematic because we cannot assume that these roles will remain stable across different languages. When looking at an automatically parsed English-Chinese parallel corpora, it was shown that 8.7% of the arguments do not preserve their semantic roles (Fung et al., 2006).

UCCA provides universal semantic structures which have a minimal set of labels. It provides a complete semantic tree which does not rely on syntactic heads and the semantic structure is grounded directly to the words in the sentence. Even though the set of UCCA labels are fairly restricted, nevertheless they allow us to determine the most important components of the graph (for example it defines centres and linkers which would be likely to carry more weight than elaborators), and we can use this to better calculate the score. We think that UCCA is the most promising representation for evaluating translation.

4.3 MT Evaluation Overview

The basic strategy for annotating the semantics of the machine translations is to step through the source sentence semantic components, looking at the translation via the word alignments, and marking on the source structure which parts have been

correctly translated. There are two kinds of components: a word or basic semantic unit, and a structural component which contains one or more sub-components.

4.3.1 Lexical nodes

Lexical nodes are usually comprised of individual words. They are the leaf nodes on the tree, the smallest meaning-bearing units of the tree. Leaf nodes can be labelled as green (correct), orange (partially correct) and red (incorrect). The traffic light system makes the marking of the lexical units as simple as possible.

- Green: The meaning of the word or phrase has been largely captured.
- Orange: the essential meaning has been captured, but some part of the translation is wrong. This could often be due to the translated word having the wrong tense, or the wrong morphology.
- Red: The essential meaning of the unit has not been captured.

4.3.2 Structural nodes

Structural nodes contain other nodes which could be either lexical or structural nodes. These nodes are also called parent nodes and they can be labelled as "Adequate" or "Bad". What we are trying to gauge for structural nodes is if the children of these nodes relate to each other in the same way in the source sentence and in the translation. There are many ways that the relationship between the children might go wrong in translation:

- One of the child nodes is missing from the aligned translation.
- An extra word or phrase has been inserted into the aligned translation.
- The components of the translation are ordered differently from the source.

If any of these changes have occurred and this damages the meaning of the aligned translation, then we mark the structural node as "Bad". However if the arrangement of the child nodes is essentially correct, except that one or more of the child nodes has themselves been translated wrongly (and so is marked orange or red), then the structural node should be annotated as "Acceptable".

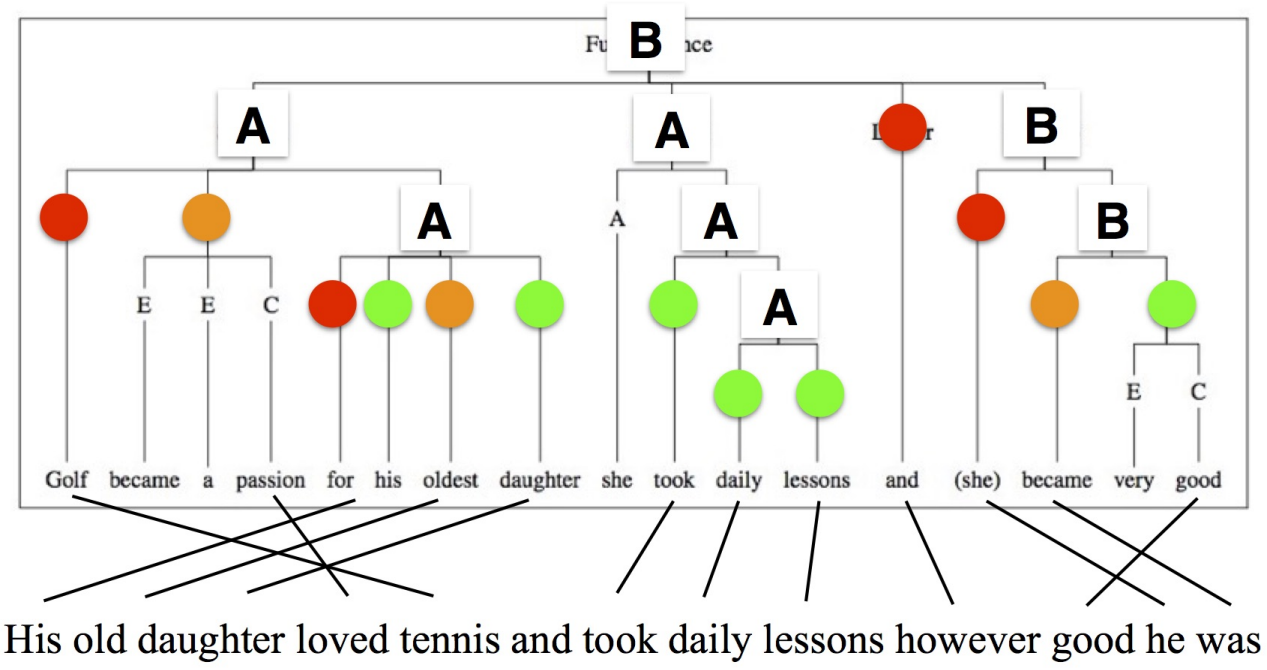

Figure 2: UCCA Tree evaluated in comparison to an aligned "translation" (an English paraphrase).

There are cases where we cannot usefully compare individual words in the source to words in the target. We often need to compare translation at the phrase level. If this is the case, then we can treat the structural node as a lexical unit, and mark it with the traffic light labels. This means that none of the children of this node will be examined in order determine the semantic score for this sentence. In Figure 2, we can see that the phrase "became a passion" has been labelled as a lexical node and it has been evaluated as "Orange", or partially correct, as the translation of this phrase "loved" only partially capture the semantics of the source.

We are separating lexical and structural evaluation in order to simplify evaluation and to localise errors to their point of origin. For this reason, if a word has been translated incorrectly, the parent node should still be correct if the number and relationship between its children are the same as in the source. In Figure 2, we can see that even though an important child of the first scene is incorrectly translated ("Golf"), and the children of are ordered differently, the scene itself is marked as "Acceptable". The translation of the scene retains the structural relationships between the children.

4.4 Machine Translation Evaluation Tool

In this section, we will describe the tool that the annotators used to perform the annotation.

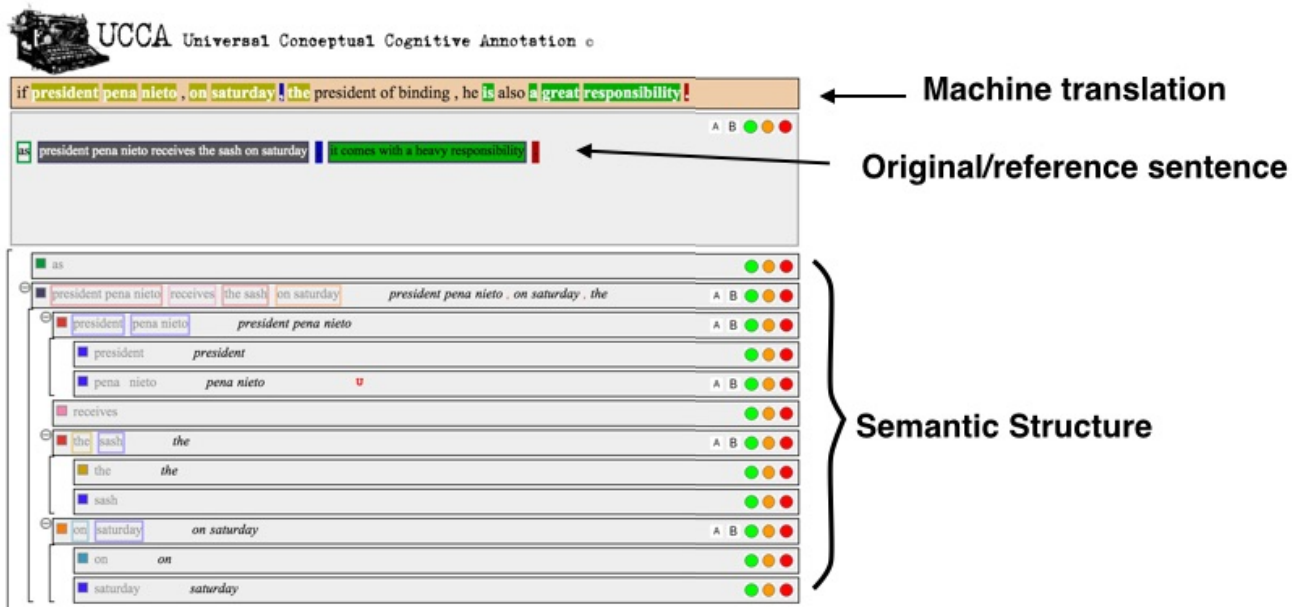

Figure 3: Machine translation evaluation tool

In Figure 3 you can see the MT evaluation tool. The sentence at the top shows the complete MT system output. Underneath the MT output is the source sentence. Underneath the source sentence we see its expandable semantic tree structure with both lexical and structural nodes. Lexical nodes only have traffic light annotation, whereas a structural node would normally be labelled "A" or "B" but could also be labelled as a lexical node, in the case where there is no word to word correspondence for the translation of its children.

The child components of a structural node are marked with different coloured rectangles. When navigating through the different source sentence nodes, one can see the relevant sections of the translation because the aligned words in the translation are highlighted in the complete sentence above. Aligned translations are also shown in black alongside the source node. If the node is aligned to a set of discontiguous words in the translation, then the unaligned words that appear in between the aligned words are shown in red. Even if these words are not directly aligned to the source node, they will likely change the meaning of the translation and must be considered when marking the translation as correct or not. The alignments are meant just as a guide. The annotator should look at the complete translation when deciding on the evaluation of a node. The translation could have content added before or after the node which changes its structure or meaning. If an extra component is prepended to a structural node, for example, it should be marked as "Bad".

4.5 Results

In this section we report the results of the evaluation. We first report some statistics which give us an overall idea of how the annotators fared. We then delve into the question of how reliable the annotations are by examining inter-annotator agreement and looking at confusion matrices. Finally we calculate UCCA semantic scores for the different language pairs.

4.5.1 Overall Statistics

In order to explore the results of our annotations, we first show some basic statistics about the task. In Table 4, we can see the total number of sentences annotated with the number of nodes. We also show the percentages of the main types of nodes. Looking at this table we see that the number of sentences varies from 324 to 351 sentences, except for Ro1 who did considerably fewer, only 230. All annotators were shown the same set of sentences, but their results were only collected if they pressed the submit button. So these missing sentences were either mistakenly not submitted, or they did not get to the end of the task. Looking at the types of nodes annotated, these seem to be distributed quite uniformly. About 35% of the nodes were considered to be structural, and 60% to be lexical. Some nodes were not evaluated, and these are called "Missing". The node may represent an English word which is not required in the target language (often an article or a pronoun). As per instructions, the annotators were expected to mark such nodes with green but some may have left the node unannotated (in which case it would be perhaps appropriate to consider the parent as a lexical node). With hindsight, we should have introduced a separate annotation category for this, and also enforced the lexical/structural distinction in the tool. "Missing" annotations are not very common, but one annotator left 6.2% of the nodes unlabelled which is a bit too high and in future experiments we will alter the tool to disallow missing nodes, unless their parent is a lexical node.

	No. Sentences	No. Nodes	%Structural	%Lexical	%Missing
De1	339	9253	36.2	62.5	1.3
Cs1	324	8794	33.4	63.3	3.3
Ro1	230	6152	36.4	62.7	0.8
Ro2	337	9228	35.6	61.4	2.9
Pl1	351	9557	34.0	59.8	6.2
Pl2	340	9303	31.0	64.6	4.4

Table 4: Number of annotated sentences and nodes and the percentages of each of the main kind of nodes

	No. Nodes	%A	%B
De1	3350	68.2	31.7
Cs1	2939	73.4	26.6
Ro1	2239	68.2	31.8
Ro2	3286	65.4	34.6
Pl1	3246	50.3	49.7
Pl2	2884	71.5	28.5

Table 5: Number of structural nodes and the percentages of each of the types of nodes

In Table 5, we can see the number of structural nodes, and their breakdown in terms of acceptable "A" or bad "B". Most annotators seem to label about 70% of the structural nodes as correct. The exception is Pl1 who seems to be following a different annotation strategy. It is possible that Pl1 did not understand the guidelines correctly. Further investigation with a Polish speaking colleague will be necessary.

	No. Nodes	%G	%O	%R
De1	5784	70.8	13.0	26.6
Cs1	5568	62.8	18.4	18.8
Ro1	3859	75.9	8.4	15.7
Ro2	5671	68.7	16.9	14.3
Pl1	5717	60.8	13.1	26.1
Pl2	6007	49.6	17.3	33.0

Table 6: Number of lexical nodes and the percentages of each of the types of nodes

In Table 6 we can see the number of lexical nodes, and their breakdown in terms of correct "G", partial "O" or incorrect "R". There is more variation here between the annotators than for the structural nodes. The percentage of correct nodes fluctuates from

49.6% to 75.9%. One expects that languages with different amounts of morphology and agglutination would report different numbers here, but even amongst annotators of the same languages the results vary. For Polish the number of correct lexical nodes varies between 49.6% and 60.8%.

4.5.2 IAA

In order to determine how reliable our human evaluation approach is, we need to examine the agreement that we find between annotators. For two of the language pairs, Romanian and Polish, we had 2 annotators. Inter-annotator agreement can be calculated using the Kappa score which is more reliable than agreement because it takes into account the probability of agreement by chance.

MT Eval Label	A	B	All
A	1096	285	1381
B	101	507	608
All	1197	792	1989

Table 7: Confusion Matrix for the structural nodes for the Romanian annotators with a Kappa of 0.58

MT Eval Label	R	O	G	All
R	2274	361	126	2761
O	92	164	37	293
G	82	76	358	516
All	2448	601	521	3570

Table 8: Confusion Matrix for the lexical nodes for the Romanian annotators with a Kappa of 0.50

In Tables 7 and 8 we see the confusion matrix of the two main types of evaluation labels for Romanian. We separate structural and lexical nodes, because there is very little disagreement between annotators about which node is a lexical and which node is structural, and so if we calculate Kappa over all the values, this might artifically inflate our IAA scores. For both structural and lexical nodes, the Kappa values of 0.58 and 0.50 respectively can be considered to display moderate agreement. If we calculate Kappa scores across all five categories, it rises to 0.69 because of the strong agreement about what is a lexical and a structural node.

MT Eval Label	A	B	All
A	1208	192	1400
B	681	574	1255
All	1889	766	2655

Table 9: Confusion Matrix for the structural nodes for the Polish annotators with a Kappa of 0.33

MT Eval Label	R	O	G	All
R	2430	444	427	3301
O	119	398	198	715
G	161	109	1110	1380
All	2710	951	1735	5396

Table 10: Confusion Matrix for the lexical nodes for the Polish annotators with a Kappa of 0.58

In Tables 9 and 10 we see the confusion matrix of the two main types of evaluation labels for Polish. Here we see similar Kappa results for the lexical nodes (0.56) but the Kappa for structural nodes drops to 0.33. This low agreement could be the result of one annotator misinterpreting the guidelines and warrants further analysis. If we calculate Kappa scores across all five categories, it rises to 0.58.

Looking beyond Kappa scores, we did some analysis of the cases where one annotator marked a word as Red and another marked it Green. For Polish the most common English words about which the annotators disagreed were: “to”, “you”, “blood” (which was aligned to “krwi”), “in” and “for”. For Romanian, the words were “you”, “to”, “the”, and “your”. Apart from “blood”, all these words are short function words which can be translated in a very different fashion in the target and this could contribute to confusion about whether they are correct or not.

Looking at the words which were marked Red most often again these were largely function words. This is partially due to the fact that they occur more frequently than other words. But words such as “bias”, “falls”, “fibrillation” and “review” also

occur frequently in this list and these are words with domain specific translations which we are possibly not capturing with the unadapted HimL models.

4.5.3 UCCA Scores

Node score This simple score reflects the percentage of correct MT evaluation nodes. A and G are correct, O counts as 50% correct and R and B count as incorrect. Nodes which are missing an evaluation are ignored.

Annotator ID	Structural	Lexical	Overall
De1	68.20	77.40	74.03
Cs1	73.35	72.00	72.46
Ro1	68.15	80.09	75.71
Ro2	65.42	77.26	72.92
Pl1	50.33	67.33	61.17
Pl2	71.53	58.26	62.56

Table 11: Node scores: percentage of correct nodes.

In Table 11 we can see the simple node scores for the different annotators. Results show that most systems are evaluated as having about 70% correct overall, with the notable exception of both Polish annotators. They gave scores to the HimL test set which were about 10% lower. English→Polish is considered to be a very challenging language pair for machine translation and so this result is not surprising. Another trend is that the annotators gave the lexical scores of at least 10% higher than the structural scores for three out of five annotators. The two remaining annotators behaved quite differently. Cs1 gave similar scores for structural and lexical nodes, but Pl2 gave much better scores for structural nodes (71.53 vs. 58.26). The differences in behaviour displayed by the two Polish annotators in this table and the low kappa reported in Table 9 brings into question the reliability of the Polish annotations.

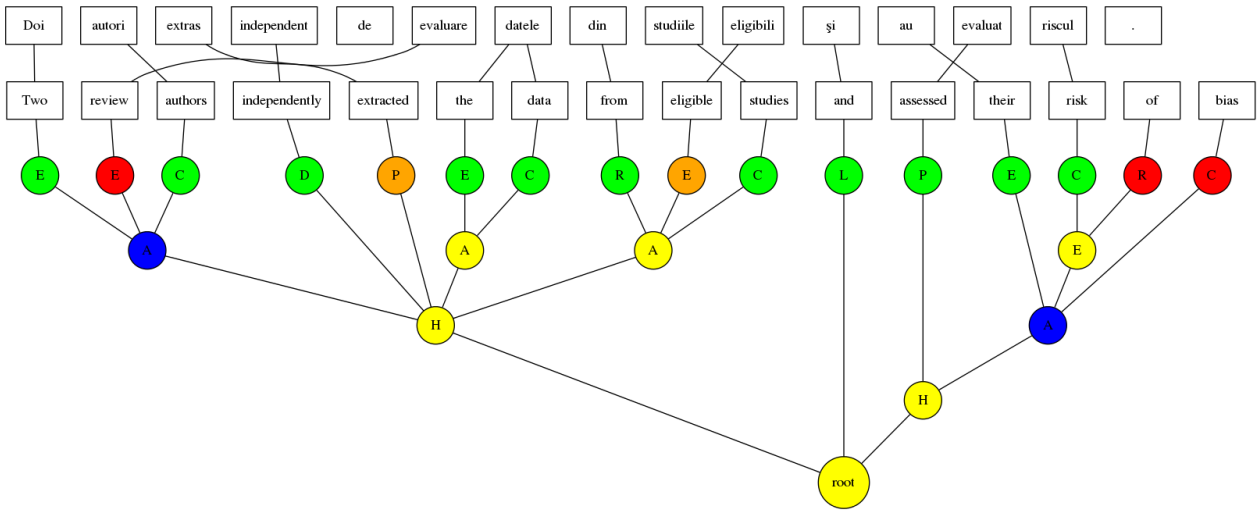

Figure 4: UCCA tree on the source labelled with the MT Eval labels and aligned to the target translation. The "Acceptable" structural nodes are yellow, the "Bad" ones are blue.

In order to visualise the UCCA evaluation results and see how the node scores are calculated, we provide an example from the experiment. Looking at Figure 4, we can see the machine translation at the top. Below this are the word alignments to the English source. Each of the English words (except punctuation) are linked to UCCA lexical nodes which have been evaluated as either correct (green), partially correct (orange) or incorrect (red). The letters inside the circles correspond to the UCCA node types which will be important in future versions of the score. Structural nodes are either acceptable (yellow) or bad (blue). So for this sentence given that we have 11 green nodes, 2 orange nodes, 6 yellow nodes and in total there are 24 nodes:

$$nodescore = \frac{11 + 0.5(2) + 6}{24} * 100 = 75$$

4.6 Discussion

In this evaluation we have proposed an evaluation which has a number of advantages:

- It shows moderate IAA agreement which means that our metric is relatively reliable. It is not directly comparable to other metrics because HMEANT papers do not report IAA and they had to rely on agreement statistics which can be misleading.
- We can reuse the UCCA structures on the source multiple times, and as this is the most time consuming part of the evaluation, this makes best use of expensive expert annotators.
- Translation annotators do not have to try to create semantic structures on erroneous machine translation output, in fact they do not need to create any semantic structures at all as this can be done previously by experts.
- Errors are isolated at the point where they occur due to the separation of structure and lexical content and there is little ambiguity about whether an error in a leaf needs to be counted multiple times as we go up the tree.
- UCCA provides a minimal yet complete semantic framework which is based on universal cognitive concepts. The structure does not rely on syntactic heads and the semantic graph is grounded directly to the words in the sentence.

The evaluation also brought up a number of issues that will need to be resolved by either better training or by improving the tool. Many of these issues were reported by the annotators:

- Isolating errors to the point where they occur is difficult to do for humans as there is a tendency to penalise the parent nodes for errors which occur in their children.
- Initially some annotators struggled to understand the distinction between lexical and structural nodes.
- They also had some difficulty evaluating the sentences without their context.
- The annotators were unsure about how to judge incorrect morphology or tense.
- The alignments were meant only as a guide and some words are unaligned or doubly aligned and this caused some confusion.
- Extraneous additions on the target might not be penalised unless they fall inside the source semantic structure.

Even though there have been a number of issues which have arisen, the UCCA evaluation has shown to be a reliable and efficient method for evaluating the accuracy of HimL translation systems. We will continue to analyse the results from this experiment and we will run an improved version of this evaluation in year two of the project.

4.7 Automatic Semantic Evaluation

The results that we have gathered in this evaluation will be used as gold data to evaluate the performance of a number of automatic metrics, to determine which of the existing metrics are most suited to evaluating accuracy. We have also initiated work on developing an automatic version of the UCCA human evaluation task. For this we need an UCCA semantic parser and data to train it, as the UCCA treebank is relatively small. We are looking at extracting further training examples from the Prague treebank t-layer. This will be described in Deliverable D5.3, due at the end of July 2016.

5 Outlook

For year two of the project, we will be pursuing the following activities:

- We will carefully analyse the performance of our models on the HimL test set to determine what improvements are necessary for the consumer health domain.
- We will evaluate research done on domain adaptation, semantics and morphology to determine which deliver performance improvements for the HimL test set. This will drive our integration efforts.
- We will be continuing development of the human semantic metric based on feedback from the first evaluation experiment.
- We will be looking at fully automating the human semantic metric.
- We will follow the evaluation plan to implement comprehensive user acceptance and impact studies.

References

Abend, O. and Rappoport, A. (2013). Universal Conceptual Cognitive Annotation (UCCA). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Ageeva, E., Tyers, F. M., Forcada, M. L., and Pérez-Ortiz, J. A. (2015). Evaluating machine translation for assimilation via a gap-filling task. In *Proceedings of EAMT 2015, The Eigtheenth Annual Conference of the European Association for Machine Translation*, pages 137–144.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of Linguistic Annotation Workshop*.

Birch, A., Haddow, B., Germann, U., Nadejde, M., Buck, C., and Koehn, P. (2013). The feasibility of HMEANT as a human MT evaluation metric. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 52–61, Sofia, Bulgaria. Association for Computational Linguistics.

Bojar, O. and Wu, D. (2012). Towards a Predicate-Argument Evaluation for MT. In *Proceedings of the Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 30–38, Jeju, Republic of Korea. Association for Computational Linguistics.

Chuchunkov, A., Tarelkin, A., and Galinskaya, I. (2014). Applying HMEANT to English-Russian Translations. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 43–50, Doha, Qatar. Association for Computational Linguistics.

Fung, P., Zhaojun, W., Yongsheng, Y., and Wu, D. (2006). Automatic learning of chinese english semantic structure mapping. In *Spoken Language Technology Workshop, 2006. IEEE*, pages 230–233. IEEE.

Lo, C.-k. and Wu, D. (2011). Structured vs. flat semantic role representations for machine translation evaluation. In *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 10–20. Association for Computational Linguistics.

Palmer, M., Gildea, D., and Xue, N. (2010). Semantic role labeling. *Synthesis Lectures on Human Language Technologies*, 3(1):1–103.

Xue, N., Bojar, O., Hajič, J., Palmer, M., Urešová, Z., and Zhang, X. (2014). Not an interlingua, but close: Comparison of english AMRs to chinese and czech. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., and Mariani, J., editors, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1765–1772, Reykjavík, Iceland. European Language Resources Association.