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

D5.3: Report on preliminary semantic evaluation metric

Author(s): Alexandra Birch, Barry Haddow, Ondřej Bojar, David Mareček, Liane Guillou

Dissemination Level: Public

Date: August, 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	August, 1st 2016
Actual date of delivery	August 1st 2016
Deliverable number	D5.3
Deliverable title	Report on preliminary semantic evaluation metric
Type	Report
Status and version	1.0
Number of pages	21
Contributing partners	UEDIN
WP leader	UEDIN
Task leader	UEDIN
Authors	Alexandra Birch, Barry Haddow, Ondřej Bojar, David Mareček, Liane Guillou
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

This document provides a summary of the work done in the first half of the HimL project with regard to developing human and automatic semantic metrics. High accuracy machine translation cannot progress without reliable metrics to measure progress.

We report on our design and implementation of a decomposable human semantic metric called HUME (Human UCCA-based MT Evaluation) in Section 2. Section 3 then reports on experiments where we automated the human evaluation process and tuned it to the human scores.

Aside from the overall semantic correctness of machine translation, we focus on one particular component: pronouns. Correct links in texts are undoubtedly very important for preserving the meaning of the text but traditional methods of MT evaluation that cover the whole sentence usually neglect this specific phenomenon. In Section 4, we explain the difficulties of evaluating pronouns and propose how a dedicated measure could be applied to HimL data.

We conclude with an outlook in Section 5.

2 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. In this project, we have developed a human evaluation method which is reliable and affordable and we apply it to the HimL, MT prototypes.

The work described in this section relates to task *T5.2: Human semantic evaluation*. In January 2016, the annual evaluation work package report *D5.2: First Evaluation Report* described the initial human evaluation experiment held in November 2015. Since then, we have performed further experiments and analysis (Birch et al., 2016). In this report, we focus on the experiments and analysis done since January.

2.1 Overview

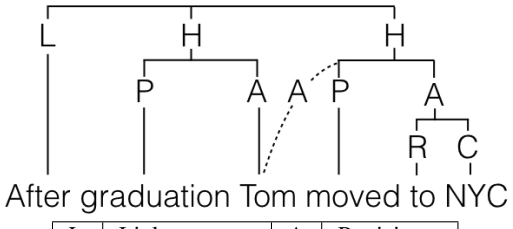
Human evaluation of machine translation normally uses sentence-level measures such as relative ranking or adequacy scales. However, these provide no insight into possible errors, and do not scale well with sentence length. We argue for a semantics-based evaluation, which captures what meaning components are retained in the MT output, providing a more fine-grained analysis of translation quality, and enables the construction and tuning of semantics-based MT. We present a novel human semantic evaluation measure, Human UCCA-based MT Evaluation (HUME), building on the UCCA semantic representation scheme. HUME covers a wider range of semantic phenomena than previous methods and does not rely on semantic annotation of the potentially garbled MT output. We experiment with four language pairs translating out of English, demonstrating HUME's broad applicability, and report good inter-annotator agreement rates and correlation with human adequacy scores.

Human judgement is the cornerstone for estimating the quality of an MT system. Nevertheless, common measures for human MT evaluation, such as adequacy and fluency judgements or the relative ranking of possible translations, are problematic in two ways. First, as the quality of translation is multi-faceted, it is difficult to quantify the quality of the entire sentence in a single number. This is indeed reflected in the diminishing inter-annotator agreement (IAA) rates of human ranking measures with the sentence length (Bojar et al., 2011). Second, a sentence-level quality score does not indicate which parts of the sentence are badly translated, and so cannot inform developers in repairing these errors.

These problems are partially addressed by measures that decompose over parts of the evaluated translation. For automatic measures, these are often words or n-grams, for manual measures some structural information is taken into account (Macháček and Bojar, 2015), or the annotators are explicitly asked to mark errors, which however suffers from even lower agreement than ranking (Lommel et al., 2014). A promising line of research decomposes metrics over semantically defined units, quantifying the similarity of the output and the reference in terms of their verb argument structure; the most notable of these measures is HMEANT (Lo and Wu, 2011).

We propose the HUME metric, a human evaluation measure that decomposes over the UCCA semantic units. UCCA (Abend and Rappoport, 2013) is an appealing candidate for semantic analysis, due to its cross-linguistic applicability, support for rapid annotation, and coverage of many fundamental semantic phenomena, such as verbal, nominal and adjectival argument structures and their inter-relations.

HUME operates by aggregating human assessments of the translation quality of individual semantic units in the source sentence. We are thus avoiding the semantic annotation of machine-generated text, which is often garbled or semantically unclear. This also allows the re-use of the source semantic annotation for measuring the quality of different translations of the same source sentence, and avoids relying on possibly suboptimal reference translations.



L	Linker	A	Participant
H	Parallel Scene	R	Relater
P	Process	C	Centre

Figure 1: Sample UCCA annotation. Leaves correspond to words and nodes to units. The dashed edge indicates that "Tom" is also a participant in the "moved to NYC" Scene. Edge labels mark UCCA categories.

After a brief review (Section 2.2), we describe HUME (Section 2.3). Our experiments with the four HimL language pairs: English to Czech, German, Polish and Romanian (Section 2.5) document HUME's inter-annotator agreement and efficiency (time of annotation). We further empirically compare HUME with direct assessment of human adequacy ratings, and conclude by discussing the differences with HMEANT (Section 2.6).

2.2 Background

MT Evaluation. Human evaluation is generally done by ranking the outputs of multiple systems e.g., in the WMT tasks (Bojar et al., 2015), or by assigning adequacy/fluency scores to each translation, a procedure recently improved by Graham et al. (2015b). However, while providing the gold standard for MT evaluation, human evaluation is not a scalable solution.

Scalability is addressed by employing automatic and semi-automatic approximations of human judgements. Commonly, such scores decompose over the sub-parts of the translation, and quantify how many of these sub-parts appear in a manually created reference translation. This decomposition allows system developers to localize the errors. The most commonly used measures decompose over n-grams or individual words, e.g., BLEU (Papineni et al., 2002), NIST (Doddington, 2002) and METEOR (Banerjee and Lavie, 2005). Another common approach is to determine the similarity between the reference and translation in terms of string edits (Snover et al., 2006). While these measures stimulated much progress in MT research by allowing the evaluation of massive-scale experiments, the focus on words and n-grams does not provide a good estimate of semantic correctness, and may favour shallow string-based MT models.

In order to address this shortcoming, more recent work quantified the similarity of the reference and translation in terms of their structure. Liu and Gildea (2005) took a syntactic approach, using dependency grammar, and Owczarzak et al. (2007) took a similar approach using lexical-functional grammar structures. Giménez and Màrquez (2007) proposed to combine multiple types of information, capturing the overlap between the translation and reference in terms of their semantic (predicate-argument structures), lexical and morphosyntactic features.

Perhaps the most notable attempt at semantic MT evaluation is MEANT and its human variant HMEANT (Lo and Wu, 2011), which quantifies the similarity between the reference and translation in terms of the overlap in their verbal argument structures and associated semantic roles. We discuss the differences between HMEANT and HUME in Section 2.6.

Semantic Representation. UCCA (Universal Conceptual Cognitive Annotation, Abend and Rappoport, 2013) is a cross-linguistically applicable, lightweight scheme for semantic annotation. Formally, an UCCA structure is a directed acyclic graph (DAG), whose leaves correspond to the words of the text. The graph's nodes, called UNITS, are either terminals or several elements jointly viewed as a single entity according to some semantic or cognitive consideration. Edges bear a category, indicating the role of the sub-unit in the structure the unit represents.

UCCA's current inventory of distinctions focuses on argument structures (adjectival, nominal, verbal and others) and relations between them. The most basic notion is the Scene, which describes a movement, an action or a state which persists in time. Each Scene contains one main relation and zero or more participants. For example, the sentence "After graduation, Tom moved to NYC" contains two Scenes, whose main relations are "graduation" and "moved". The participant "Tom" is a part of both Scenes, while "NYC" only of the latter (Figure 1). Further categories account for inter-scene relations and the sub-structures of participants and relations.

The use of UCCA for semantic MT evaluation measure is motivated by two main reasons. First, UCCA's set of categories can be reliably annotated by non-experts after as little as two hours of training (Marinotti, 2014). Second, UCCA is cross-linguistically

applicable, seeking to represent what is shared between languages by building on linguistic typological theory (Dixon, 2010b,a, 2012). Its cross-linguistic applicability has so far been tested in annotations of English, French, German and Czech.

The Abstract Meaning Representation (AMR) (Banarescu et al., 2013) project shares UCCA's motivation for defining a more complete semantic annotation. However, using AMR is not optimal for defining a decomposition of a sentence into semantic units as it does not ground its semantic symbols in the text, and thus does not provide clear decomposition of the sentence into sub-units. Also, AMR is more fine-grained than UCCA and consequently harder to annotate. Other approaches represent semantic structures as bi-lexical dependencies (Sgall et al., 1986; Hajič et al., 2012; Oepen and Lønning, 2006), which are indeed grounded in the text, but are less suitable for MT evaluation as they require knowledge of formal linguistics for their annotation.

2.3 The HUME Measure

2.3.1 Annotation Procedure

This section summarises the manual annotation procedure used to compute the HUME measure. We denote the source sentence as s and the translation as t. The procedure involves two manual steps: (1) UCCA-annotating s, (2) human judgements as to the translation quality of each semantic unit of s relative to t, where units are defined according to the UCCA annotation. UCCA annotation is performed once for every source sentence, irrespective of the number of its translations we wish to evaluate

UCCA Annotation. We begin by creating UCCA annotations for the source sentence, following the UCCA guidelines[1]. A UCCA annotation for a sentence s is a labeled DAG G, whose leaves are the words of s. For every node in G, we define its *yield* to be its leaf descendants. The semantic units for s according to G are the yields of nodes in G.

Translation Evaluation. HUME annotation is done by traversing the semantic units of the source sentence, which correspond to the arguments and relations expressed in the text, and marking the extent to which they have been correctly translated. HUME aggregates the judgements of the users into a composite score, which reflects the overall extent to which the semantic content of s is preserved in t.

Annotation of the semantic units requires first deciding whether a unit is *structural*, i.e., has meaning-bearing sub-units also in the target language, or *atomic*. In most cases, atomic units correspond to individual words, but they may also correspond to unanalyzable multi-word expressions. When a multi-word unit is labeled as atomic, its sub-units' annotations are ignored in the evaluation.

Atomic units can be labelled as "Green" (correct), "Orange" (partially correct) and "Red" (incorrect). Green means that the meaning of the word or phrase has been largely preserved. Orange means that the essential meaning of the unit has been preserved, but some part of the translation is wrong. This is often be due to the translated word having the the wrong inflection, in a way that impacts little on the understandability of the sentence. Red means that the essential meaning of the unit has not been captured.

Structural units have sub-units (children in the UCCA graph), which are themselves atomic or structural. Structural units are labeled as "Adequate" or "Bad", meaning that the relation between the sub-units went wrong[2]. We will use the example "man bites dog" to illustrate typical examples of why a structural node should be labelled as "Bad": incorrect ordering ("dog bites man"), deletion ("man bites") and insertion ("man bites biscuit dog").

HUME labels reflect adequacy, rather than fluency judgements. Specifically, annotators are instructed to label a unit as Adequate if its translation is understandable and preserves the meaning of the source unit, even if its fluency is impaired.

Figure 2 presents an example of a HUME annotation, where the translation is in English for ease of comprehension. When evaluating "to NYC" the annotator looks at the translation and sees the word "stateside". This word captures the whole phrase and so we mark this non-leaf node with an atomic label. Here we choose Orange since it approximately captures the meaning in this context. The ability to mark non-leaves with atomic labels allows the annotator to account for translations which only correspond at the phrase level. Another feature highlighted in this example is that by separating structural and atomic units, we are able to define where an error occurs, and localise the error to its point of origin. The linker "After" is translated incorrectly as "by" which changes the meaning of the entire sentence. This error is captured at the atomic level, and it is labelled Red. The sentence still contains two scenes and a linker and therefore we mark the root node as structurally correct, Adequate.

[1] All UCCA-related resources can be found here: http://www.cs.huji.ac.il/~oabend/ucca.html

[2] Three labels are used with atomic units, as opposed to two labels with structural units, as atomic units are more susceptible to slight errors.

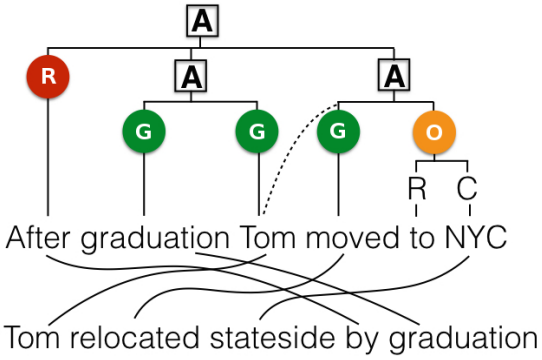


Figure 2: HUME annotation of an UCCA tree with a word aligned example translation shown below. Atomic units are labelled using traffic lights (Red, Orange, Green) and structural units are marked A or B.

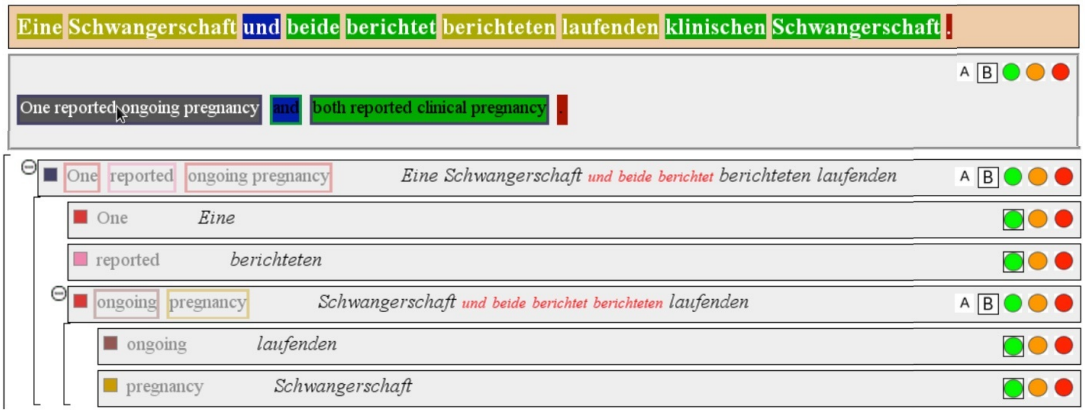


Figure 3: The HUME annotation tool. The top orange box contains the translation. The source sentence is directly below it, followed by the tree of the source semantic units. Alignments between the source and translation are in italics and unaligned intervening words are in red (see text).

2.3.2 Composite Score

We proceed to detailing how judgements on the semantic units of the source are aggregated into a composite score. We start by taking a very simple approach and compute an accuracy score. Let $Green(s,t)$, $Adequate(s,t)$ and $Orange(s,t)$ be the number of Green, Adequate and Orange units, respectively. Let $Units(s)$ be the number of units marked with any of the labels. Then HUME's composite score is:

$$HUME(s,t) = \frac{Green(s,t) + Adequate(s,t) + 0.5 \cdot Orange(s,t)}{Units(s)}$$

2.4 Annotation Interface

Figure 3 shows the HUME annotation interface. The user is asked to select a label for each source semantic unit, by clicking the "A", "B", Green, Orange, or Red buttons to the right of the unit's box. Units with multiple parents (as with "Tom" in Figure 2) are displayed twice, once under each of their parents, but are only annotatable in one of their instances, to avoid double counting.

The interface presents, for each unit, the translation segment aligned with it. This allows the user, especially in long sentences, to focus her attention on the parts most likely to be relevant for her judgement. As the alignments are automatically derived, and therefore noisy, the annotator is instructed to treat the aligned text is a cue, but to ignore the alignment if it is misleading, and instead make a judgement according to the full translation. Concretely, let s be a source sentence, t a translation, and $A \subset 2^s \times 2^t$ a many-to-many word alignment. If u is a semantic unit in s, whose yield is $yld(u)$, we define the aligned text in t to be $\bigcup_{(x_s,x_t)\in A \wedge x_s \cap yld(u) \neq \emptyset} x_t$.

Where the aligned text is discontinuous in t, words between the left and right boundaries which are not contained in it (intervening words) are presented in a smaller red font. Intervening words are likely to change the meaning of the translation of u, and

		cs	de	pl	ro
#Sentences	Annot. 1	324	339	351	230
	Annot. 2	205	104	340	337
#Units	Annot. 1	8794	9253	9557	6152
	Annot. 2	5553	2906	9303	9228

Table 1: HUME-annotated #sentences and #units.

	cs	de	pl	ro
Annot. 1	255	140	138	96
Annot. 2	*	162	229	207

Table 2: Median annotation times per sentence, in seconds. *: no timing information is available, as this was a collection of annotators, working in parallel.

thus should be attended to when considering whether the translation is correct or not.

For example, in Figure 3, "ongoing pregnancy" is translated to "Schwangerschaft ... laufenden" (lit. "pregnancy ... ongoing"). This alone seems acceptable but the interleaving words in red notify the annotator to check the whole translation, in which the meaning of the expression is not preserved. The annotator should thus mark this structural node as Bad.

2.5 Experiments

In order to validate the HUME metric, we ran an annotation experiment with one source language (English), and four HimL languages (Czech, German, Polish and Romanian), using text from the public health domain. Semantically accurate translation is paramount in this domain, which makes it particularly suitable for semantic MT evaluation. HUME is both evaluated in terms of its consistency (inter-annotator agreement), efficiency (time of annotation) and validity (through a comparison with crowd-sourced adequacy judgements).

2.5.1 Datasets and Translation Systems

For each of the four language pairs under consideration we built phrase-based MT systems using Moses (Koehn et al., 2007). These were trained on large parallel data sets extracted from OPUS (Tiedemann, 2009), and the data sets released for the WMT14 medical translation task (Bojar et al., 2014), giving between 45 and 85 million sentences of training data, depending on language pair. These translation systems were used to translate texts derived from both NHS 24[3] and Cochrane[4] into the four languages. NHS 24 is a public body providing healthcare and health-service related information in Scotland, Cochrane is an international NGO which provides independent systematic reviews on health-related research. NHS 24 texts come from the "Health A-Z" section in the NHS Inform website, and Cochrane texts come from their plain language summaries and abstracts.

2.5.2 HUME Annotation Statistics

The source sentences are all in English, and their UCCA annotation was performed by four computational linguists and one linguist. For the annotation of the MT output, we recruited two annotators for each of German, Romanian and Polish and one main annotator for Czech. For Czech IAA, several further annotators worked on a small number of sentences each. We treat these further annotators as one annotator, resulting in two annotators for each language pair. The annotators were all native speakers of the respective target languages and fluent in English.

Table 1 shows the total number of sentences and units annotated by each annotator. Not all units in all sentences were annotated, often due to the annotator accidentally missing a node.

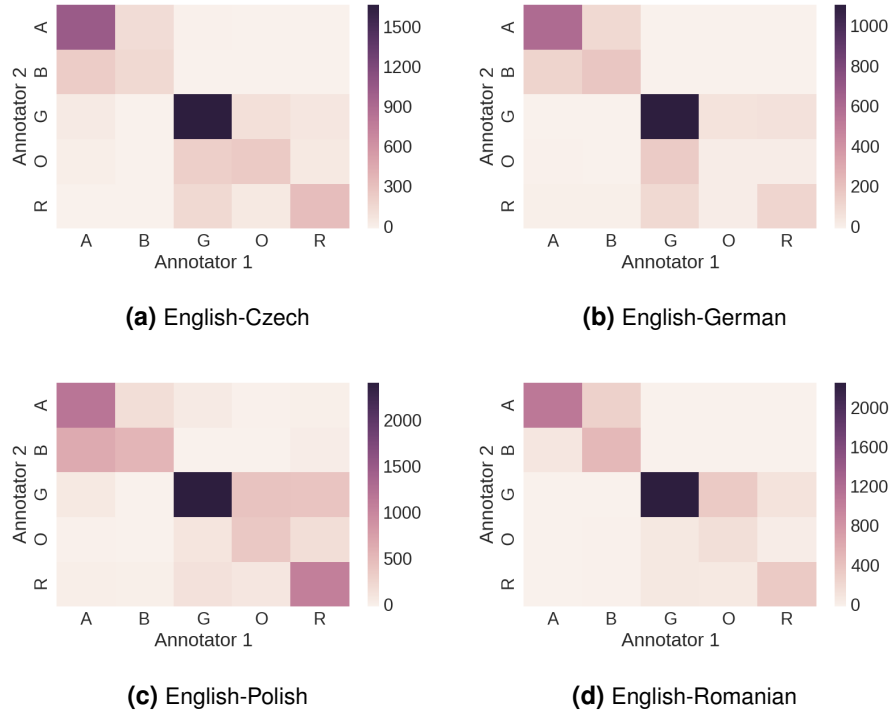
Efficiency. We estimate the annotation time using the timestamps provided by the annotation tool, which are recorded whenever an annotated sentence is submitted. Annotators are not able to re-open a sentence once submitted. To estimate the annotation time, we compute the time difference between successive sentences, and discard outlying times since we assume annotation was

[3] http://www.nhs24.com/
[4] http://www.cochrane.org/

	cs	de	pl	ro
Sentences	181	102	334	217
All units	4686	2793	8384	5604
Kappa	0.64	0.61	0.58	0.69
Atomic units	2982	1724	5386	3570
Kappa	0.54	0.29	0.54	0.50
Structural units	1602	1040	2655	1989
Kappa	0.31	0.44	0.33	0.58

Table 3: IAA for the multiply-annotated units, measured by Cohen's Kappa.

not continuous. From inspection of histograms of annotation times, we set the upper threshold at 500 seconds. Median annotation times are presented in Table 2, indicating that the annotation of a sentence takes around 2–4 minutes, with some variation between annotators.



(a) English-Czech **(b)** English-German

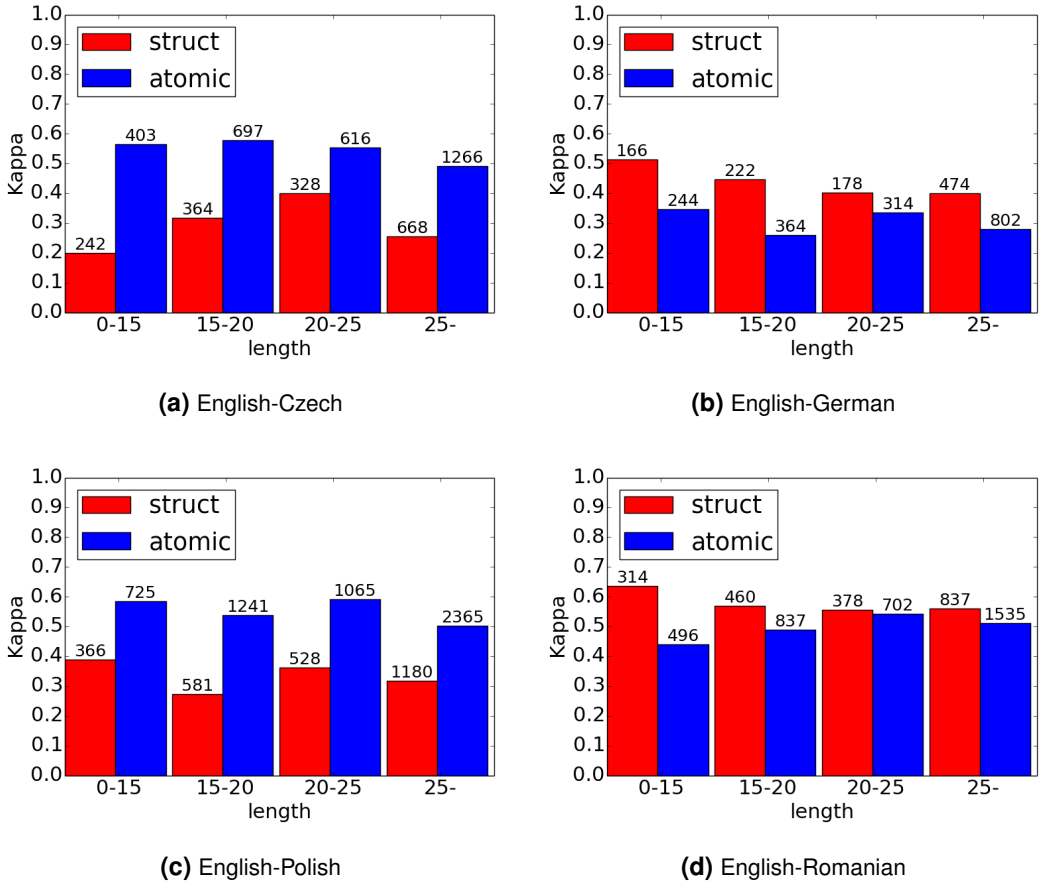
(c) English-Polish **(d)** English-Romanian

Figure 4: Confusion matrices for each language pair.

Inter-Annotator Agreement. In order to assess the consistency of the annotation, we measure the Inter-Annotator Agreement (IAA) using Cohen's Kappa on the multiply-annotated units. Table 3 reports the number of units which have two annotations from different annotators and the corresponding Kappas. We report the overall Kappa, as well as separate Kappas on atomic units (annotated as Red, Orange or Green) and structural units (annotated as Adequate or Bad). As expected and confirmed by confusion matrices in Figure 4, there is generally little confusion between the two types of units.

To assess HUME reliability for long sentences, we binned the sentences according to length and measured Kappa on each bin (Figure 5). We see no discernible reduction of IAA with sentence length. Also, from Table 3 the overall IAA is similar for all languages, showing good agreement (0.6–0.7). However, there are differences observed when we break down by node type. Specifically, we see a contrast between Czech and Polish, where the IAA is higher for atomic than for structural units, and German and Romanian, where the reverse is true. We also observe low IAA (around 0.3) in the cases of German atomic units, and Polish and Czech structural units.

Looking more closely at the areas of disagreement, we see that for the Polish structural units, the proportion of As was quite different between the two annotators (53% vs. 71%), whereas for other languages the annotators agree in the proportions. We believe that this was because one of the Polish annotators did not fully understand the guidelines for structural units, and



(a) English-Czech **(b)** English-German

(c) English-Polish **(d)** English-Romanian

Figure 5: Kappa versus sentence length for structural and atomic units. (Node counts in bins on top of each bar.)

percolated errors up the tree, creating more Bs. For German atomic and Czech structural units, where Kappa is also around 0.3, the proportion of such units being marked as "correct" is relatively high, meaning that the class distribution is more skewed, so the expected agreement used in the Kappa calculation is high, lowering Kappa. Finally we note some evidence of domain-specific disagreements, for instance the German MT system normally translated "review" (as in "systematic review" – a frequent term in the Cochrane texts) as "überprüfung", which one annotator marked correct, and the other (a Cochrane employee) as incorrect.

2.5.3 Comparison with Direct Assessment

Recent research (Graham et al., 2015b,a; Graham, 2015) has proposed a new approach for collecting accuracy ratings, direct assessment (DA). Statistical interpretation of a large number of crowd-sourced adequacy judgements for each candidate translation on a fine-grained scale of 0 to 100 results in reliable aggregate scores, that correlate very strongly with one another.

We attempted to follow Graham et al. (2015b) but struggled to get enough crowd-sourced judgements for our target languages. We ended up with 10 adequacy judgements on most of the HUME annotated translations for German and Romanian but insufficient data for Czech and Polish. We see this as a severe practical limitation of DA.

Figure 6 plots the HUME score for each sentence against its DA score. HUME and Direct Assessment scores correlate reasonably well. The Pearson correlation for en-ro (en-de) is 0.70 (0.58), or 0.78 (0.74) if only doubly HUME-annotated points are considered. This confirms that HUME is consistent with an accepted human evaluation method, despite the differences in their conception. While DA is a valuable tool, HUME has two advantages: it returns fine-grained semantic information about the quality of translations and it only requires very few annotators. Direct assessment returns a single opaque score, and (as also noted by Graham et al.) requires a large crowd which may not be available or reliable.

Figure 7 presents an analysis of HUME's correlations with DA by HUME unit type, an analysis enabled by HUME's semantic decomposition. For both target languages, correlation is highest in the 'all' case, supporting our claim for the value of aggregating over a wide range of semantic phenomena. Some types of nodes predict the DA scores better than others. HUME scores

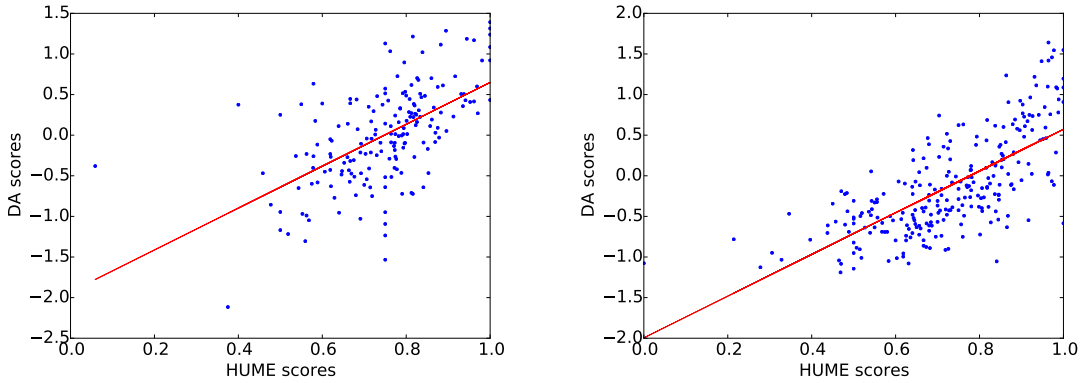


Figure 6: HUME vs DA scores. DA score have been standardised for each crowdsourcing annotator and averaged across exactly 10 annotators. HUME scores are averaged where there were two annotations.

on As correlate more strongly with DA than scores on Scene Main Relations (P+S). Center nodes (C) are also more correlated than elaborator nodes (E), which is expected given that Centers are defined to be more semantically dominant. Future work will construct an aggregate HUME score which weights the different node types according to their semantic relevance.

2.6 Comparison with HMEANT

We discuss the main differences between HUME and HMEANT, a human MT evaluation metric that measures the overlap between the translation a reference in terms of their SRL annotations.

Verbal Structures Only? HMEANT focuses on verbal argument structures, ignoring other pervasive phenomena such as non-verbal predicates and inter-clause relations. Consider the following example:

Source	a coronary angioplasty may not be technically possible
Transl.	eine koronare Angioplastie kann nicht technisch möglich
Gloss	a coronary angioplasty can not technically possible

The German translation is largely correct, except that the main verb “sein” (“be”) is omitted. While this may be interpreted as a minor error, HMEANT will assign the sentence a very low score, as it failed to translate the main verb. Conversely, HMEANT does not penalize errors such as tense or negation flip in a correctly aligned predicate.

We conducted an analysis of the English UCCA Wikipedia corpus (5324 sentences) in order to assess the pervasiveness of three phenomena that are not well supported by HMEANT.[5] First, copula clauses are treated in HMEANT simply as instances of the main verb “be”, which generally does not convey the meaning of these clauses. They appear in 21.7% of the sentences, according to conservative estimates that only consider non-auxiliary instances of “be”. Second, nominal argument structures, ignored by HMEANT, are in fact highly pervasive, appearing in 48.7% of the sentences. Third, linkers that express inter-relations between clauses (mainly discourse markers and conjunctions) appear in 56% of the sentences, but are again ignored by HMEANT. As noted in our experiments, linkers are sometimes omitted in translation, but these omissions are not taken into consideration by HMEANT.

We are not aware of any empirical argument suggesting that verb argument structures, taken alone, capture the crux of the sentence semantics. Moreover, relying only on verbal argument structures is less stable across paraphrases and translations, as a verbal argument structure may be translated to a nominal or adjectival argument structure (e.g., “after graduation” may be translated into “after he graduated”). This may lead to an unduly low HMEANT score, as the verb in one structure has nothing to align to in the other. On the other hand, UCCA has been shown to be reasonably stable in an English-French corpus study (Sulem et al., 2015).

[5] Argument structures and linkers are explicitly marked in UCCA. Non-auxiliary instances of “be” and nouns are identified using the NLTK standard tagger. Nominal argument structures are here Scenes whose main relation is headed by a noun.

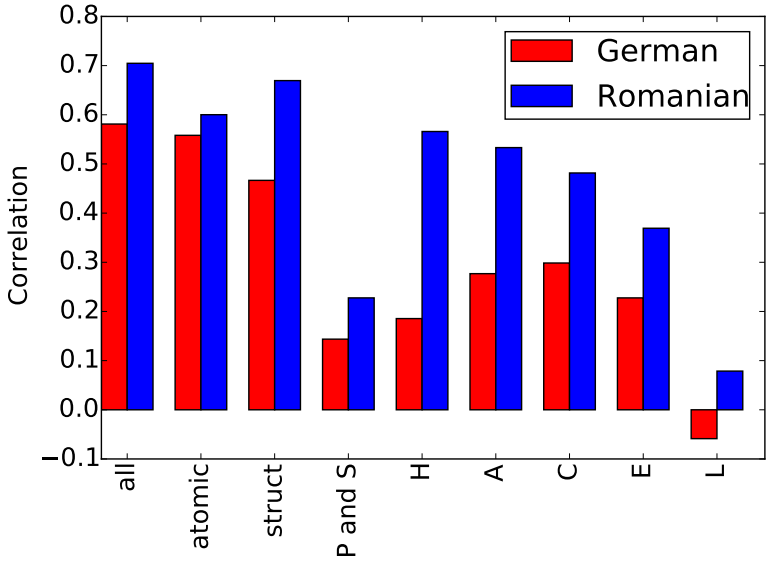


Figure 7: Pearson correlation of HUME vs. DA scores for en-ro and en-de. Each bar represents a correlation between DA and an aggregate HUME score based on a sub-set of the units (#nodes for the en-de/en-ro setting in brackets): all units ('all', 8624/10885), atomic ('atomic', 5417/6888) and structural units ('struct', 3207/3997), and units by UCCA categories: Scene main relations (i.e, Process and State units; 'P and S', 954/1178), Parallel Scenes ('H', 656/784), Participants ('A', 1348/1746), Centres ('C', 1904/2474), elaborators ('E', 1608/2031) and linkers ('L', 261/315).

We note that some of these issues were already observed in previous applications of HMEANT to languages other than English. See Birch et al. (2013) for German, Bojar and Wu (2012) for Czech and Chuchunkov et al. (2014) for Russian.

One Structure or Two. HUME only annotates the source, while HMEANT relies on two independently constructed structural annotations, one for the reference and one for the translation. Not annotating the translation is appealing as it is often impossible to assign a semantic structure to a low quality translation. On the other hand, HUME may be artificially boosting the perceived understandability of the translation by allowing access to the source.

Alignment. In HMEANT, the alignment between the reference and translation structures is a key part of the manual annotation. If the alignment cannot be created, the translation is heavily penalized. Bojar and Wu (2012) and Chuchunkov et al. (2014) argue that the structures of the reference and of an accurate translation may still diverge, for instance due to a different interpretation of a PP-attachment, or the verb having an additional modifier in one of the structures. It would be desirable to allow modifications to the SRL annotations at the alignment stage, to avoid unduly penalizing such spurious divergences. The same issue is noted by Lo and Wu (2014): the IAA on SRL dropped from 90% to 61% when the two aligned structures were from two different annotators. HUME uses automatic (word-level) alignment, which only serves as a cue for directing the attention of the annotators. The user is expected to mentally correct the alignment as needed, thus circumventing this difficulty.

Monolingual vs. Bilingual Evaluation. HUME diverges from HMEANT and from shallower measures like BLEU, in not requiring a reference. Instead, it compares the source directly with the output translation. This requires the employment of bilingual annotators, but has the benefit of avoiding using a reference, which is never uniquely defined, and may thus lead to unjustly low scores where the translation is a paraphrase of the reference.

Error Localisation. In HMEANT, an error in a child node often results in the parent node being penalised as well. This makes it harder to quantify the true scale of the original error, as its effect gets propagated up the tree. In HUME, errors are localised as much as possible to where they occur, by the separation of atomic and structural units, which supports a more accurate aggregation of the translation errors to a composite score.

2.7 Summary

We have introduced HUME, a human semantic MT evaluation measure which addresses a wide range of semantic phenomena. We have shown that it can be reliably and efficiently annotated in multiple languages, and that annotation quality is robust to

sentence length. Comparison to direct assessments further support HUME's validity. We believe that HUME allows for a more fine-grained analysis of translation quality, and will be a useful tool to guide the development of a more semantically aware approach to MT.

3 Automatic Semantic Evaluation

The annotations that we gathered in the evaluation described in Section 2 are used as gold data in the search for and development of automatic semantic metric.

In Section 3.1, we evaluate to what extent existing MT metric correlate with our HUME scores.

Fully automating HUME is a rather complex task. We would need an UCCA semantic parser and data to train it, as the UCCA treebank is relatively small. We would also need to automate the test that our bilingual annotators did, i.e. checking if individual semantic components of the source sentence were preserved in the translation. In the long term, we are considering methods that would allow us to reach UCCA annotations via existing (and automated) representations, such as the tectogrammatical layer developed in Prague or AMR (see Section 2.2 discussing the differences between UCCA and AMR). We are also looking at extracting further training examples from the tectogrammatical annotations available in Prague dependency treebanks (Hajič et al., 2012; Hajič et al., 2006).

For the time being, we take the pragmatic approach and search for methods that replicate *final HUME scores* well, not necessarily following its structure in any way. When HUME composite score (Section 2.3.2) becomes more complex, the importance of the automatic metrics following UCCA structure will grow.

In Section 3.1, we search for a good correlate of HUME among existing segment-level MT metrics. In Section 3.2, we develop our own metric based on the tectogrammatical representation. Since this representation is readily available only for Czech, we propose and evaluate an approximation of it based on Universal Dependencies in Section 3.3.

3.1 Correlation of HUME with Standard MT Metrics

There are many automatic metrics that report scores at the level of individual sentences (segment-level). If we find a metric that correlates very well with HUME at the segment level, this would provide us with a good proxy for an automatic metric based directly on HUME.

We measured Pearson correlation of a few variants of the MT metrics NIST, BLEU and chrF (all are based on matching word or character n-grams with the reference translation) against the HUME scores. We also included our HUME golden data as one of the subtasks of the Shared Metrics Task[6] (Bojar et al., 2016) to enlarge the set of examined metrics. The correlations are listed in Tables 4 and 5, respectively. Some metrics appear in both evaluations with slightly different scores. This is possible due to marginal differences in metric implementations used in the references; the underlying set of sentences and golden HUME scores were identical.

The evaluation in the metrics task included also an assessment of confidence. Metrics not significantly outperformed by any other in a given language pair can be thus highlighted in bold in Table 5 and they constitute the winners of the shared task.

We see that the best-performing automatic metrics reach correlation levels of between .4 and .6. This does not seem particularly high, but there is no directly comparable dataset. The metrics task (Bojar et al., 2016) shows segment-level correlations for an extended set metrics with a different type of manual quality assessment ("Direct Assessment", DA, see the metrics task paper), with correlations in the range of .6 to .7 for the best performing metrics. Note that the set of languages is also different so the numbers cannot be directly compared.

Interestingly, the best correlating metrics are based on simple *character-level* matches with the reference, which is arguably quite different from the semantic basis of HUME.

3.2 Using Deep Syntactic Features for Czech

To obtain an automatic metric which correlates with HUME better than metrics examined in Section 3.1 and also to experiment with representations closer to the semantics of the sentence, we developed our own metric. The metric is a simple linear regression combining several features extracted from the source and reference. Some of the features rely on automatic tectogrammatical annotation. This annotation is available only for Czech, so this section experiments solely with this single language. See Section 3.3 below for an approximation applicable to all HimL languages.

[6] http://www.statmt.org/wmt16/metrics-task/

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
BLEU	0.361	0.404	0.314	0.538
BLEU cased	0.350	0.406	0.316	0.535
chrF3	0.540	0.511	0.419	0.638
chrF1	0.505	0.497	0.428	0.608

Table 4: Pearson correlations of different metrics against HUME evaluated for this report.

Metric	en-cs	en-de	en-pl	en-ro
CHRF3	**.544**	**.480**	.413	**.639**
CHRF2	.537	**.479**	**.417**	.634
BEER	.516	**.480**	**.435**	.620
CHRF1	.506	**.467**	**.427**	.611
MPEDA	.468	**.478**	**.425**	.595
WORDF3	.413	.425	.383	.587
WORDF2	.408	.424	.383	.583
WORDF1	.392	.415	.381	.569
SENTBLEU	.349	.377	.328	.550

Table 5: Pearson correlation of segment-level metric scores taking part in the HUME subtask of WMT16 metrics task, reproduced from Bojar et al. (2016).

3.2.1 Word Alignment between Translation and Reference

Our automated metric relies on automatic alignment between the translation candidate and the reference translation. The easiest way of obtaining word alignments is to run GIZA++ (Och and Ney, 2000) on the set of sentence pairs. GIZA was designed to align documents in two languages and it can obviously also align documents in a single language, although it does not benefit in any way from the fact that many words are identical in the aligned sentences. GIZA works well if the input corpus is sufficiently large, to allow for extraction of reliable word co-occurrence statistics.

While the HUME test set alone is too small, we have a corpus of paraphrases for Czech (Bojar et al., 2013). We thus run GIZA++ on all possible paraphrase combinations together with the reference-translation pairs we need to align and then extract alignments only for the sentences of interest.

3.2.2 Parsing up to the Tectogrammatical Level

We use Treex[7] framework to do the tagging, parsing and tectogrammatical annotation. Tectogrammatical annotation of sentence is a dependency tree, in which only content words are represented by nodes.[8] The main label of the node is a tectogrammatical lemma – mostly the same as the morphological lemma, sometimes combined with a function word in case it changes its meaning. Other function words and grammatical features of the words are expressed by other attributes of the tectogrammatical node. The main attributes are:

- **tectogrammatical lemma (t-lemma)**: the lexical value of the node,
- **functor**: semantic values of syntactic dependency relations. They express the functions of individual modifications in the sentence, e.g. ACT (Actor), PAT (Patient), ADDR (Addressee), LOC (Location), MANN (Manner),
- **sempos**: semantic part of speech: n (noun), adj (adjective), v (verb), or adv (adverbial),
- **formeme**: morphosyntactic form of the node. The formeme includes for example prepositions and cases of the nouns.
- **grammatemes**: tectogrammatical counterparts of morphological categories, such as number, gender, person, negation, modality, aspect, etc.

An example of a pair of tectogrammatical trees is provided in Section 8.

[7] `http://ufal.mff.cuni.cz/treex`

[8] Arguably, this is not yet fully semantic representation as UCCA but it is as close as we can get automatically.

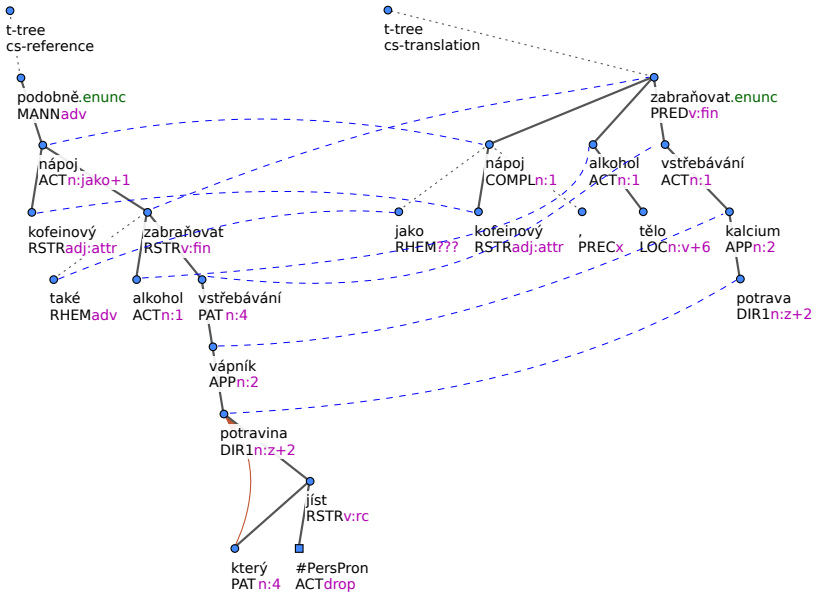


Figure 8: Example of aligned tectogrammatical trees of reference and translation.

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
linear regression	0.625
linear regression + feature selection	0.659

Table 6: Czech Deep-syntactic features and their correlation against HUME.

3.2.3 Scores Expressing the Ratio of Matching Attributes

Given the word- (or node-) alignment links between tectogrammatical annotations of the translation and reference sentences, we can count a percentage of links where individual attributes agree, e.g. the number of pairs of tectogrammatical nodes that have the same tectogrammatical lemma. These scores capture only a portion of what the tectogrammatical annotations offer, for instance, we they do not consider the structure of the trees at all.

For the time being, we take these scores as individual features and use them in a combined model, see the next section.

3.2.4 Training Linear Regression on the Golden HUME Scores

We collected all the scores based on matching of tectogrammatical attributes, added BLEU scores (on forms and lemmas), and chrF scores (3-grams and 6-grams) and trained a linear regression model to obtain a mix of features that fits best the HUME scores. Since the amount of annotated data available is low, we use the jackknife strategy:

- We split the annotated data into ten parts.
- For each tenth, we train the regression on all the rest data and apply it to this tenth.

By this procedure, we obtain automatically assigned scores for all sentences in the data. The correlation coefficients are shown in Table 6, along with the individual features.

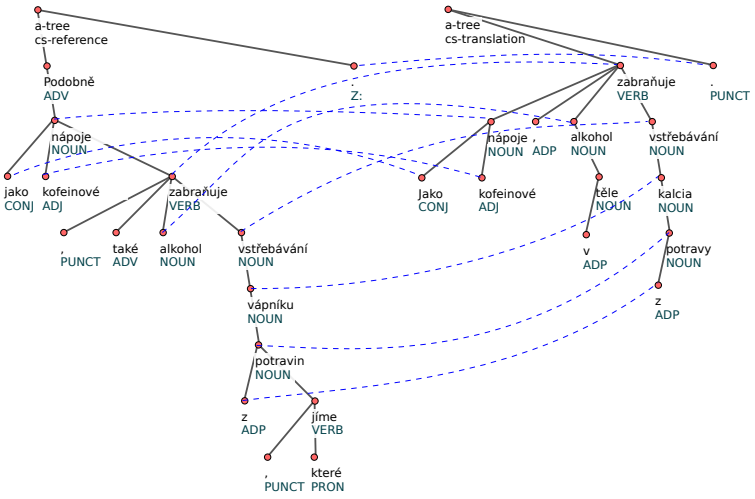


Figure 9: Example of aligned dependency trees (only universal POS tags are shown within the nodes).

In addition to the regression using all the features, we also did a feature selection, in which we manually chose only features with a positive impact on the overall correlation score. For instance, we found that the BLEU scores can be easily omitted without worsening the correlation. Conversely, the chrF scores are very valuable and omiting them would lower the correlation significantly.

We see that chrF3 alone performs reasonably well (Pearson of .54), and we know from Table 5 that it was also the winner of the HUME track of the metrics task. If we combine it with a selected subset our features, we are able to achieve the correlation of up to .659.

3.3 Deeper Features for All HimL Languages

We have seen that deep-syntactic features help to train an automatic metric with higher correlation for Czech. Even though we have no similar tools for German, Polish, and Romanian so far, we try to extract similar features for them. Unlike the language-specific approach in Section 3.2, this approach is language universal as much as possible.

3.3.1 Universal Parsing.

We use Universal Dependencies (UD) by Nivre et al. (2016), a collection of treebanks (40 languages in the current version 1.3) in common annotation style, where all our testing languages are present. For syntactic analysis, we use UDPipe by Straka et al. (2016) – tokenizer, tagger, and parser in one tool, which is trained on UD. The UD tagset consists of 17 POS tags, the big advantage is that the tagset is the same for all the languages and therefore we can easily extract e.g. content words, prepositional phrases, etc.

3.3.2 Monolingual Alignment

We have no corpus of paraphrases for German, Polish, and Romanian, so we used a simple monolingual aligner based on word similarities and relative positions in the sentence. First, we compute scores for all possible alignment connections between tokens of the reference and translated sentence.

$$score(i,j) = w_1 JaroWinkler(W_i^t, W_j^r) + w_2 \mathrm{I}(T_i^t = T_j^r) + w_3(1 - |(i/len(t) - j/len(r)|),$$

where $JaroWinkler(W_i^t, W_j^r)$ defines similarity between the given words, $\mathrm{I}(T_i^t = T_j^r)$ is a binary indicator testing the identity of POS tags, and $(1 - |(i/len(t) - j/len(r)|)$ tells us how close are the two words according to their relative positions in the sentences. The weights were set manually to $w_1 = 8$, $w_2 = 3$, and $w_3 = 3$. When we have the scores, we can simply produce unidirectional alignments (i.e. find the best token in the translation for each token in the reference and vice versa) and then symmetrize them to create intersection and union alignments.

Figure 9 provides an illustration of Czech UD trees, aligned at the level of node.

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
chrF3	0.540	0.511	0.419	0.638
chrF1	0.505	0.497	0.428	0.608
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
linear regression	0.604	0.525	0.453	0.656

Table 7: Pearson correlations of different metrics against HUME.

3.3.3 Extracting Features

We distinguish content words from function ones by the POS tag. The tags for nouns (NOUN, PROPN), verbs (VERB), adjectives (ADJ), and adverbs (ADV) correspond more or less to content words. Then there are pronouns (PRON), symbols (SYM), and other (X), which may be sometimes content words as well, but we do not count them. The rest of POS tags represent function words.

Now, using the alignment links and the content words, we can 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. Our features can thus check for instance the number of aligned words with matching number or tense.

3.3.4 Regression and Results

We compute all the scores proposed in the previous section on the four languages and test the correlation with HUME. German UD annotation does not contain lemmas, so some scores for German could not be computed. The results are shown in Table 7. Similarly as in Section 3.2, we trained a linear regression on all the features together with chrF scores.

3.4 Discussion and Future Work

We proposed several variations of a metric based on matching features between words in the reference and translated sentences. Even though no one alone outperformed the chrF3 metric in terms of correlation with HUME score, the linear regression over all of them (including chrF3) trained on the manually annotated data reached much better correlations for all the four languages.

Our experiments indicate that tectogrammatical annotation of Czech helped to get better correlation scores (0.659) than the simplified version using using the UD annotation only (0.604).

In future work, we plan to automatically simulate directly the HUME metric. We will create a parser (or a tree convertor) for the source sentences to get the UCCA structures and automatically assign scores to individual tokens and phrases.

4 Semi-automatic Pronoun Evaluation

The evaluation of pronoun translation poses a particular challenge to MT researchers. Automatic metrics, such as BLEU (Papineni et al., 2002), which are typically used in MT evaluation follow the assumption that overlap of MT output with a human-generated reference translation may be used as a proxy for correctness. In the case of anaphoric pronouns, which corefer with a noun phrase (the *antecedent*), this assumption breaks down. If the pronoun's antecedent is translated in a way that differs from the reference translation, a different pronoun may be required. It may in fact be wrong to use a pronoun that matches the one in the reference. Furthermore, comparison with a reference translation may result in valid alternative translations being marked as incorrect.

PROTEST (Guillou and Hardmeier, 2016) comprises a test suite and a semi-automatic evaluation method for pronoun translation. It aims to reduce manual evaluation effort, and to address the problems of incorrect automatic evaluation of valid alternative

	anaphoric								event	pleonastic	addressee reference		
	it				*they*			*it/they*	*it*	*it*	*you*		
	intra		inter		intra	inter	sing.	group			generic	deictic	
	subj.	non-subj.	subj.	non-subj.								sing.	plural
Examples	*25*	*15*	*25*	*5*	*25*	*25*	*15*	*10*	*30*	*30*	*20*	*15*	*10*
Baseline	8	1	**11**	1	12	**12**	8	6	**15**	**18**	**13**	**9**	9
auto-postEDIt	**10**	**6**	6	**2**	**13**	11	8	**7**	6	11	12	8	**10**

Table 8: Matches per category for the DiscoMT 2015 shared task baseline and a participating system

translations and absolute reliance on the reference translation. The test suite includes a hand selected set of pronouns categorised according to their *function* and other features. These categories represent some of the different problems that MT systems face when translating pronouns. The pronouns in the existing test suite were selected from the *DiscoMT2015.test* set (Hardmeier et al., 2016) using annotations that follow the ParCor guidelines (Guillou et al., 2014). Under this annotation scheme, pronouns are labelled according to one of eight functions: anaphoric, cataphoric, pleonastic, event reference, speaker reference, addressee reference, extra-textual reference, or "other" function. Additional features are recorded for some pronoun functions. For example, anaphoric pronouns are linked to their antecedents, and are marked as being inter- or intra-sentential.

To construct a new test suite for the HimL test data, we would first need to annotate the test set according to the ParCor guidelines. The annotations are then used to group pronouns together into categories. For small test sets, we may wish to include all pronouns, but for larger sets we may wish to select a subset of the pronouns so as to ensure that the manual annotation effort is manageable. The selection of pronouns may also be used to balance the test set such that all categories are represented, or that the sets are not biased towards a particular expected target language pronoun token, based on the reference translation (e.g. a particular gender in the case of anaphoric pronouns). The HimL test sets comprise a subset of sentences extracted from complete documents. For the purpose of annotation, the information needed to disambiguate the function of ambiguous pronouns may lie outside the current sentence. Therefore, it is necessary to annotate complete documents.

The automatic evaluation method developed for PROTEST compares, for each pronoun in the test suite, the translation in the MT output with that in the reference. Results are provided in terms of "matches". In the case of anaphoric pronouns, the translation of both the pronoun and the head of its antecedent in the MT output must match those in the reference. For all other pronoun functions, only the translations of the pronoun must match. Those translations for which these conditions do not hold are termed "mismaches" and are referred for manual evaluation to determine whether they represent valid alternative translations or incorrect translations. An example of the results obtained with PROTEST on the *DiscoMT2015.test* set can be seen in Table 8.

The following steps outline the application of the automatic evaluation method to the HimL test suite. The evaluation method relies on word alignments between the source text and the MT output, and between the source text and the reference translation. The former will need to be obtained from the decoder, and the latter computed using a tool such as GIZA++. The source text, its translation and the word alignments are then input to PROTEST, for both the source-MT and source-reference pairs. Automatic evaluation then computes a list of matches and mismatches.

There are two proposed use cases for PROTEST. The first is to compare the evaluation scores for two systems where one system is an extension of the other, thereby checking that the quality of pronoun translation has not been degraded. For HimL, we would compare the pronoun translations by the Y1 and Y2 systems. The second use case is to provide a complete evaluation of all pronouns. Here, manual evaluation of the "mismatches" is used to complement the automatic evaluation. The manual evaluation of pronouns may be conducted using the graphical pronoun analysis tool provided for PROTEST (Hardmeier and Guillou, 2016). Manual evaluation for HimL will require the availability of annotators fluent in English plus German, Czech, Polish or Romanian.

Annotation of the documents from which the HimL test set was selected, has already begun, according to the ParCor guidelines provided for the written text genre[9]. The other steps outlined in this section form the plan for work to be carried out once the annotation work is complete.

5 Outlook

The HimL description of work sets out a timeline for the development of human and automatic semantic metrics for machine translation. The development phase of these metrics is coming to a close, but as we have plans to deploy these metrics in evaluating HimL systems annually, some refinement of these methods will occur.

[9] ParCor guidelines also exist for the spoken genre, in which exhibits some differences in pronoun use when compared to the written genre. These guidelines have been used in the annotation of TED Talks.

We plan to run another human evaluation using the HUME metric later this year, to compare the HimL Y1 and Y2 systems. This will involve improvements to the tool, and to the annotation guidelines. We will also investigate different ways of weighting the individual components of the HUME metric to better reflect their importance in the sentence.

Further development of the automatic version of our semantic metric and of the method of pronoun evaluation is also to be expected.

All these improvements will be described in the upcoming deliverable *D5.5 Report on integrated semantic evaluation metric*, and also in the yearly reports on MT evaluation (*D5.4* and *D5.6*).

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.

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 and Interoperability with Discourse*, pages 178–186.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, MI, USA. Association for Computational Linguistics.

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

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., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. (2014). Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Huck, M., Hokamp, C., Koehn, P., Logacheva, V., Monz, C., Negri, M., Post, M., Scarton, C., Specia, L., and Turchi, M. (2015). Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal. Association for Computational Linguistics.

Bojar, O., Ercegovčević, M., Popel, M., and Zaidan, O. F. (2011). A grain of salt for the WMT manual evaluation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland.

Bojar, O., Graham, Y., , and Stanojević, A. K. M. (2016). Results of the WMT16 Metrics Shared Task . In *Proceedings of the First Conference on Machine Translation*, Berlin, Germany. Association for Computational Linguistics.

Bojar, O., Macháček, M., Tamchyna, A., and Zeman, D. (2013). Scratching the Surface of Possible Translations. In *Proc. of TSD 2013*, Lecture Notes in Artificial Intelligence, Berlin / Heidelberg. Západočeská univerzita v Plzni, Springer Verlag.

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.

Dixon, R. M. (2010a). *Basic Linguistic Theory: Grammatical Topics*, volume 2. Oxford University Press.

Dixon, R. M. (2010b). *Basic Linguistic Theory: Methodology*, volume 1. Oxford University Press.

Dixon, R. M. (2012). *Basic Linguistic Theory: Further Grammatical Topics*, volume 3. Oxford University Press.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145, San Diego, CA, USA. Morgan Kaufmann Publishers Inc.

Giménez, J. and Màrquez, L. (2007). Linguistic features for automatic evaluation of heterogenous mt systems. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 256–264.

Graham, Y. (2015). Improving evaluation of machine translation quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1804–1813, Beijing, China. Association for Computational Linguistics.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2015a). Can machine translation systems be evaluated by the crowd alone? *Natural Language Engineering*, pages 1–28.

Graham, Y., Mathur, N., and Baldwin, T. (2015b). Accurate evaluation of segment-level machine translation metrics. In *Proc. of NAACL-HLT*, pages 1183–1191.

Guillou, L. and Hardmeier, C. (2016). Protest: A test suite for evaluating pronouns in machine translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 636–643. European Language Resources Association (ELRA).

Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B. (2014). Parcor 1.0: A parallel pronoun-coreference corpus to support statistical mt. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3191–3198, Reykjavik, Iceland. European Language Resources Association (ELRA).

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In *Proceedings of the Eighth International Language Resources and Evaluation Conference (LREC'12)*, pages 3153–3160, Istanbul, Turkey. ELRA, European Language Resources Association.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., and Ševčíková Razímová, M. (2006). Prague Dependency Treebank 2.0. LDC2006T01, ISBN: 1-58563-370-4.

Hardmeier, C. and Guillou, L. (2016). A graphical pronoun analysis tool for the PROTEST pronoun evaluation test suite. *Baltic Journal of Modern Computing. Special Issue: Proceedings of EAMT 2016*, pages 318–330.

Hardmeier, C., Tiedemann, J., Nakov, P., Stymne, S., and Versely, Y. (2016). DiscoMT 2015 shared task on pronoun translation. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. `http://hdl.handle.net/11372/LRT-1611`.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Liu, D. and Gildea, D. (2005). Syntactic features for evaluation of machine translation. In *ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

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.

Lo, C.-K. and Wu, D. (2014). On the Reliability and Inter-Annotator Agreement of Human Semantic MT Evaluation via HMEANT. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Lommel, A. R., Popovic, M., and Burchardt, A. (2014). Assessing Inter-Annotator Agreement for Translation Error Annotation. In *MTE: Workshop on Automatic and Manual Metrics for Operational Translation Evaluation*. LREC.

Macháček, M. and Bojar, O. (2015). Evaluating Machine Translation Quality Using Short Segments Annotations. *The Prague Bulletin of Mathematical Linguistics*, 103:85–110.

Marinotti, P. (2014). Measuring semantic preservation in machine translation with HCOMET: human cognitive metric for evaluating translation. Master's thesis, University of Edinburgh.

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. (2016). 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.

Oepen, S. and Lønning, J. T. (2006). Discriminant-based mrs banking. In *Proceedings of LREC*, pages 1250–1255.

Owczarzak, K., van Genabith, J., and Way, A. (2007). Evaluating machine translation with lfg dependencies. *Machine Translation*, 21(2):95–119.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA. Association for Computational Linguistics.

Sgall, P., Hajičová, E., and Panevová, J. (1986). *The Meaning of the Sentence and Its Semantic and Pragmatic Aspects*. Academia/Reidel Publishing Company, Prague, Czech Republic/Dordrecht, Netherlands.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, pages 223–231.

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).

Sulem, E., Abend, O., and Rappoport, A. (2015). Conceptual annotations preserve structure across translations: A French-English case study. In *ACL 2015 Workshop on Semantics-Driven Statistical Machine Translation (S2MT)*, pages 11–22.

Tiedemann, J. (2009). News from OPUS – a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248, Borovets, Bulgaria. John Benjamins.