HCOMET

Measuring Semantic Preservation in Machine Translation with HCOMET: Human Cognitive Metric for Evaluating Translation

Pedro Marinotti

Master of Science
Cognitive Science and Natural Language Processing
School of Informatics
University of Edinburgh
2014

Abstract

Human ranking of machine translation output is a commonly used method for comparing different innovations in machine translation research. Theoretically simple, the comparison of multiple translations is, in effect, cognitively complex, requiring judges to balance the weight of different types of translation errors in the context of the whole sentence. This cognitive complexity is made evident through low intra- and inter-annotator agreements, which call into question the reliability of such ranking metrics. HMEANT (Lo and Wu, 2011) attempted to decrease the complexity of ranking by dividing sentences into smaller semantic units whose translation alignments were more objective, rendering the task cognitively simpler. However, HMEANT does not discern how these semantic units are related and relies heavily on language-dependent verb frames – a significant problem for a translation metric. This project defines a new set of human metrics focusing on HCOMET (Human COgnitive Metric for Evaluating Translation). HCOMET, attempting to overcome the limitations of HMEANT, employed a new cognitively-informed annotation scheme (Abend and Rappoport, 2013) and new scoring guidelines. While the inter-annotator agreement did not surpass that of HMEANT, the conceptual framework of HCOMET allows for a much more detailed analysis of semantic adequacy in machine translation.

Acknowledgements

This project would not have been possible without the dedicated supervision of my advisor Alexandra Birch. Her knowledge, support, and engagement inspired me and allowed this project to flourish. I want to express my gratitude to Omri Abend without whom HCOMET would have met many more technical and theoretical difficulties. I would also like to thank Barry Haddow for his insight into past experiments and methods of analysis. Finally, I am also grateful for the participation of our various annotators, whose time and effort made this whole experiment possible.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Pedro Marinotti

(*Pedro Marinotti*)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Research in machine translation (MT) has evolved from early publications dating as far back as the 1960s (ALPAC, 1966) to the modern ACL Workshops on Statistical Machine Translation (WMT) (Bojar et al., 2013). In this span of more than 50 years, innovation has led to the proliferation of rule-based, syntax-based, human-assisted, and phrase-based machine translation[1] (Bojar et al., 2013; Koehn, 2009). In the midst of this variability, hundreds[2] of automatic metrics have been proposed for system optimization, but ultimately, all rely on human metrics as the ground truth for ranking and comparison (Bojar et al., 2013). Despite the critical centrality of human metrics, no such metric has been deemed an adequate measure of translation quality (Birch et al., 2013).

While it is clear that, due to a lack of discrimination and content-awareness, automatic metrics do not provide sufficiently accurate translation evaluations, modern human metrics also fail to fulfill research requirements because of low reliability and robustness, evidenced by insufficient levels of inter-annotator agreement (Papineni et al., 2002; Bojar et al., 2013). This project defines a new human metric, HCOMET[3], that, to better suit the needs of the machine translation community, incorporates cognitive and linguistic knowledge into the evaluation of semantic preservation in machine translation. Furthermore, two other metrics, LEAF[4] and SCENE[5], were also developed to

[1] Just in 2013,143 different systems were compared in the WMT (Bojar et al., 2013).

[2] Research teams proposed 55 different methods of automatic quality estimation just in the 2013 WMT alone (Bojar et al., 2013).

[3] HCOMET: Human COgnitive Metric for Evaluating Translation

[4] LEAF: Leaf Equivalence Assessment Function

[5] SCENE: Scene Complete Equivalence Numeric Evaluation

allow a more discriminative approach to the evaluation of machine translation.

1.2 Context

1.2.1 Why Automatic Metrics Require Human Metrics

Metrics such as BLEU are fast, inexpensive, and consistent[6] (Papineni et al., 2002). They may take into account *n*-grams, word-similarities, multiple reference translations, reordering, and may attempt to capture local lexico-semantic features; however, they are still proxies for translation quality and, consequently, face two issues: One, translations are produced for human use; fundamentally, a translation's quality should be determined by humans based on their ability to understand the machine translation. Two, "human understanding of the real world" enables judges to differentiate between errors in translation that introduce important ambiguity and those that do not (Dorr et al., 2011). Consequently, automatic metrics must be optimized to predict human evaluation results, further requiring the use of a human metric from which the gold standard MT evaluations can be derived.

1.2.2 Human Evaluation Strategies

Given the inadequacies of automatic methods, human metrics are the default ground truth for MT evaluation. The 2013 WMT, for example, measured the quality of automatic translation metrics by comparing their results to those of human judges (Bojar et al., 2013). Yet, such a comparison is still imperfect since even after 50 years of research, inter-annotator agreement (IAA) of human evaluation on "MT quality is surprisingly low" (Turian et al., 2003). Even intra-annotator agreement is only "moderate" when evaluating the quality of full sentences (Callison-Burch et al., 2008).

Low IAA can be ascribed to two simple reasons: One, "humans will have different opinions". Two, "judges from different backgrounds tend to weight characteristics of a translation such as syntax 'errors'[...]differently" (Dorr et al., 2011). These reasons are why human metrics of semantic adequacy were "not objective or fine-grained enough to provide a useful numeric representation of MT output quality in many situations" (Przybocki et al., 2006).

[6] NIST, WER, PER, GTM, TER, CDER, METEOR as well as BLEU are currently popular automatic metrics used in MT research and described in (Dorr et al., 2011).

To help solve these problems, research on semantic MT evaluation, of which this project is part, has given rise to more objective human metrics. By breaking down sentences into small semantic units, the evaluation of whether each unit has been translated properly becomes a cognitively simpler task, resulting in a more consistent, reliable metric.

1.2.3 Semantic MT Evaluation

Semantic MT evaluation refers to the measurement of how much meaning has been preserved through the translation from source to target language. This measurement is called semantic adequacy and has been historically contrasted to measures in fluency, which assesses how well translations follow the syntactic requirements and stylistic conventions of the target language, regardless of semantic content. However, research has found that “human annotators are not able to separate these two evaluation dimensions easily”, and that a single numerical metric may better represent translation quality (Birch et al., 2013).

HMEANT

One metric of semantic preservation, HMEANT (Lo and Wu, 2011) was tested by Birch et al. (2013) in an attempt to quantify the semantic adequacy of MT output by aligning smaller atomic semantic units within translated sentences.

Based on frame semantics and semantic role labeling (SRL), HMEANT uses human annotators to align the “who, what, whom, when, where, why” and “how” of each reference translation and its machine translation in order to calculate how much meaning was maintained. Sentences were broken down into single verb-headed frames, each carrying arguments that fit the aforementioned roles. Alignment of these frames allowed for the calculation of a more objective and quantitative evaluation score. The quantitative nature of HMEANT also allowed for the analysis of inter-annotator agreement. While the semantic role label IAAs in the HMEANT experiment were “disappointing” and did not provide a more reliable or robust metric, a protocol and baseline were created to test the viability of new human alignment based evaluation schemes (Birch et al., 2013). This project inherited many aspects of the HMEANT protocol and compared its results to the HMEANT baseline in order to provide a clear and comparable analysis of HCOMET.

1.3 Overview

In order to properly analyze HCOMET as a machine translation metric for semantic evaluation, this study was divided into six steps represented in Figure 1.1. As this firgure indicates, these same steps discribe the chapters presented in this study.

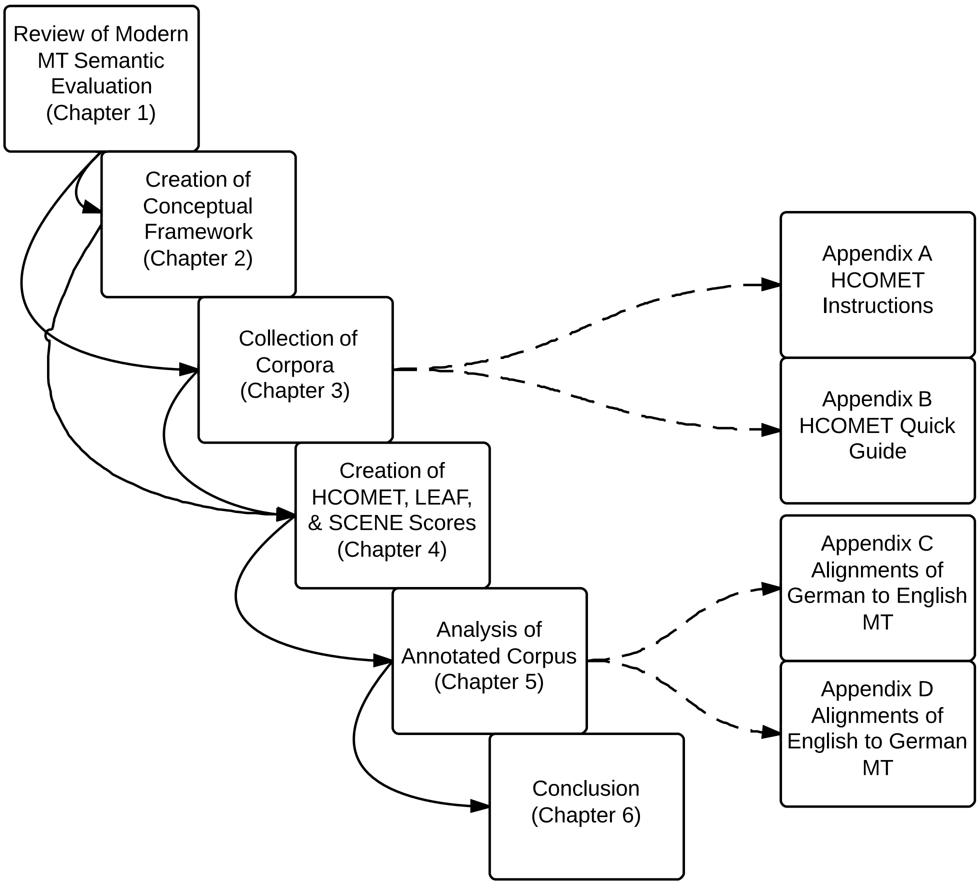

Review of Modern MT Semantic Evaluation (Chapter 1)
Creation of Conceptual Framework (Chapter 2)
Collection of Corpora (Chapter 3)
Creation of HCOMET, LEAF, & SCENE Scores (Chapter 4)
Analysis of Annotated Corpus (Chapter 5)
Conclusion (Chapter 6)
Appendix A HCOMET Instructions
Appendix B HCOMET Quick Guide
Appendix C Alignments of German to English MT
Appendix D Alignments of English to German MT

Figure 1.1: Overview of HCOMET Experiment

Chapter 2

Conceptual Framework

2.1 Universal Conceptual Cognitive Annotation (UCCA) For HCOMET

The cognitive-linguistic core of HCOMET is derived from its annotation scheme, UCCA, Universal Conceptual Cognitive Annotation. Originally created as a open framework for semantic representation, UCCA is based on Cognitive Grammar (Langacker, 2008) and Basic Linguistic Theory (BLT) (Dixon, 2010a,b, 2012). As an annotation scheme, it is meant to be "portable across domains and languages" and relatively "insensitive to meaning-preserving syntactic variation" (Abend and Rappoport, 2013). These properties make it a viable candidate for MT evaluation and theoretically account for several important shortcomings of HMEANT.

Although HMEANT attempted to be syntactically independent through its use of semantic role labels (SRLs), its infrastructure was inherently syntax-based as it required arguments to be placed in verb-headed frames. UCCA, on the other hand, does not base its atomic units of semantic representation upon any syntactic construction. The annotation scheme maps language to a "collection of [*s*]*cenes*", where each *scene* describes "some movement or action, or a temporarily persistent state [. . . or] a schematized event which refers to many events by highlighting a common meaning component" (Abend and Rappoport, 2013).

Sentences such as "(a) 'John took a shower' and (b) 'John showered"' would not align within the HMEANT infrastructure but would be annotated as the same *scene* in UCCA. Moving away from syntax, or even previously defined SRLs, UCCA attempts to take into account the cognitive representation of processes and states by annotating

Table 2.1: The UCCA categories used for annotation and their explanations, adapted from Abend and Rappoport (2013)

Abb.	Category	Short Definition
		Scenes
H	Parallel Scene	The basic top-level unit of a Scene. It is used when Scenes are not participants or Elaborators.
		Scene Elements
P	Process	The main relation of a Scene that evolves in time (usually an action or movement).
S	State	The main relation of a Scene that does not evolve in time.
A	Participant	A participant in a Scene in a broad sense (including locations, abstract entities and Scenes serving as arguments).
D	Adverbial	Used alter the semantic content of their Scenes (they including modals, manners, and sub-events).
T	Time	Used to specify the time in which the Scene or some part of it happened.
G	Ground	Used to link the scene to the speech event as opposed to another scene.
		Elements of Non-Scene Units
C	Center	Necessary for the conceptualization of the parent unit.
E	Elaborator	A non-Scene relation which applies to a single Center.
N	Connector	A non-Scene relation which applies to two or more Centers, highlighting a common feature.
R	Relator	All other types of non-Scene relations. Two varieties: (1) Rs that relate a C to some super-ordinate relation, and (2) Rs that relate two Cs pertaining to different aspects of the parent unit.
		Inter-Scene Relations
L	Linker	A relation between two or more Scenes (e.g., when, if, in order to).
		Other
F	Function	Does not introduce a relation or participant. Required by the structural pattern in which it appears.

text with labels based on cross-linguistic semantic similarity. Table 2.1 describes the UCCA labels used in this project based on on Abend and Rappoport (2013).

Using the idea of a Scene as described, we are able to break down a fairly complex sentence into smaller units of meaning as shown in Table 2.2. Each of the scenes can then be further annotated as shown in Table 2.3. The comments below the annotations demonstrate how UCCA is able to depict the semantic structure more adequately than syntax-based or verb-based annotation schemes.

In this project, the process of breaking down sentences into scenes and then annotating their internal composition was used to more objectively define the semantics of reference and machine translations so that HCOMET could better compare them.

Table 2.2: Example of UCCA Scenes in practice. The example sentence was adapted from a longer example in Abend and Rappoport (2013)

Full Sentence
"Golf became a passion for his oldest daughter: she took daily lessons and became very good."
Scenes and Linkers
[Golf became a passion for his oldest daughter] [and] [she took daily lessons] [became very good]

Table 2.3: Example of UCCA annotations in practice. The example sentence was adapted from a longer example in Abend and Rappoport (2013)

Full Sentence	
"Golf became a passion for his oldest daughter: she took daily lessons and became very good."	
Annotation of Individual Scenes	
Scene 1:	Golf became a passion for his oldest daughter
Annotation:	Golf$_A$ [became$_E$ a$_E$ passion$_C$]$_P$ [for$_R$ his$_E$ oldest$_E$ daughter$_C$]$_A$
Comments:	Although this is the longest scene, UCCA breaks it down into a simple scene with one process and two participants. The copula, *become* is not the head of the process even though it is a verb; the semantic center is *passion*.
Scene 2:	she took daily lessons
Annotation:	she$_A$ [took$_F$ [daily$_E$ lessons$_C$]$_C$]$_P$
Comments:	Notice that although HMEANT would make *took* the head of this scene, UCCA takes into account that it is merely a light verb and makes *lessons* the center of the process.
Linker:	and
Annotation:	and$_L$
Comments:	This linker connects Scenes 2 and 3. Since there are no subordinate scenes in this example, we do not need more complex linkage relationships.
Scene 3:	became very good
Annotation:	(she$_A$) [became$_E$ [very$_E$ good$_C$]$_C$]$_S$
Comments:	This scene contained an implicit unit, in this case, an implicit participant, which was made explicit during the annotation process.

2.2 HCOMET

2.2.1 Adaptations

As explained in the previous section, UCCA attempts to describe the internal semantic structure of a sentence or larger unit of text. HCOMET employs UCCA to compare two different translations in order to calculate the translation quality. However UCCA cannot be used directly for this comparison. It includes features that do not necessarily transfer into a comparative context. Certain changes needed to be made in order to provide a better tool with which two translations could be semantically compared.

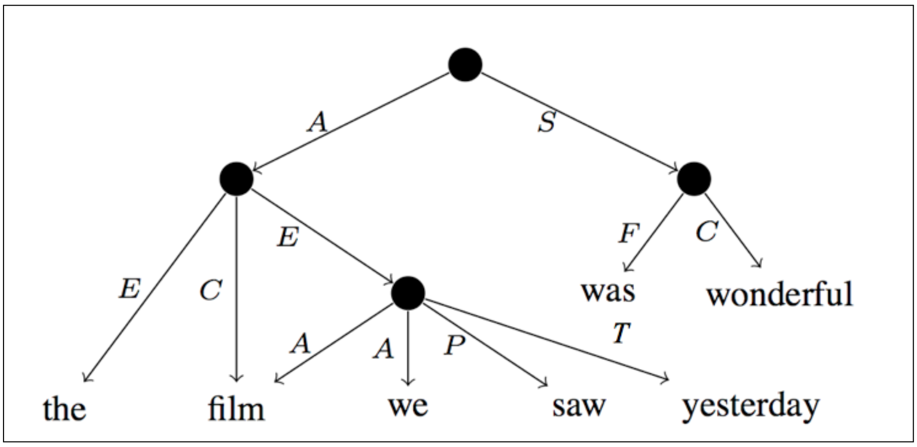
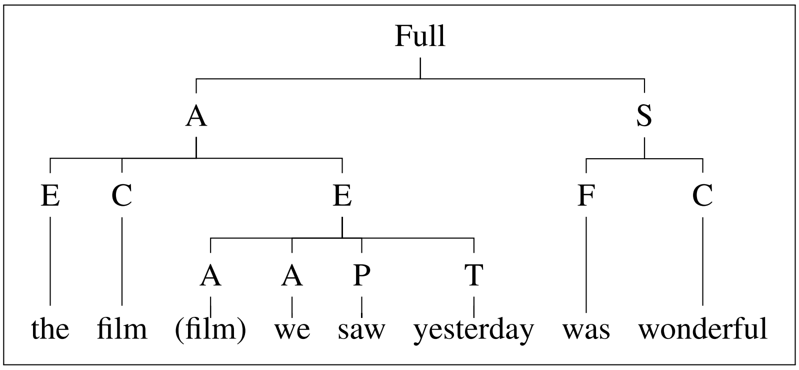


Figure 2.1: Internal Representation of UCCA Nodes and Edges from Abend and Rappoport (2013)

One such UCCA feature that required adaptation was the internal representation of nodes, specifically remote nodes. As described in Abend and Rappoport (2013), the type relation of a node is not stored in the node itself but in the edges leading to it as shown in Figure 2.1. The single node '*film*' fulfills two roles, as the *center* of '*the film we saw yesterday*' and as a *remote participant* of '*we watched yesterday*'.

A problem arises when we compare the sentence in Figure 2.1 to '*We watched a film yesterday. The film was wonderful*'. In this case, even though the semantics are exactly the same, the construction would be different since there would be two '*film*' nodes. To rectify this problem, HCOMET does not allow a node to have more than one parent. Instead, it duplicates the child node as in Figure 2.2. As a result, HCOMET's internal representation is in the form of a single rooted tree as opposed to UCCA, which only needs to be a directed acyclic graph. Figure 2.3 demonstrates how the longer sentence from Table 2.3 is represented in HCOMET.

Figure 2.2: HCOMET's Tree Structure and Typed Nodes



Another feature which required changes was the UCCA Function type (F). These units do not reflect the semantics of the sentence but are required by syntax[1]. Therefore, when comparing the semantic construction of two different translations, their presence or lack thereof, does not matter and should not be included in calculations of semantic similarity. Thus, HCOMET removes them from consideration before a user is asked to align two translations.

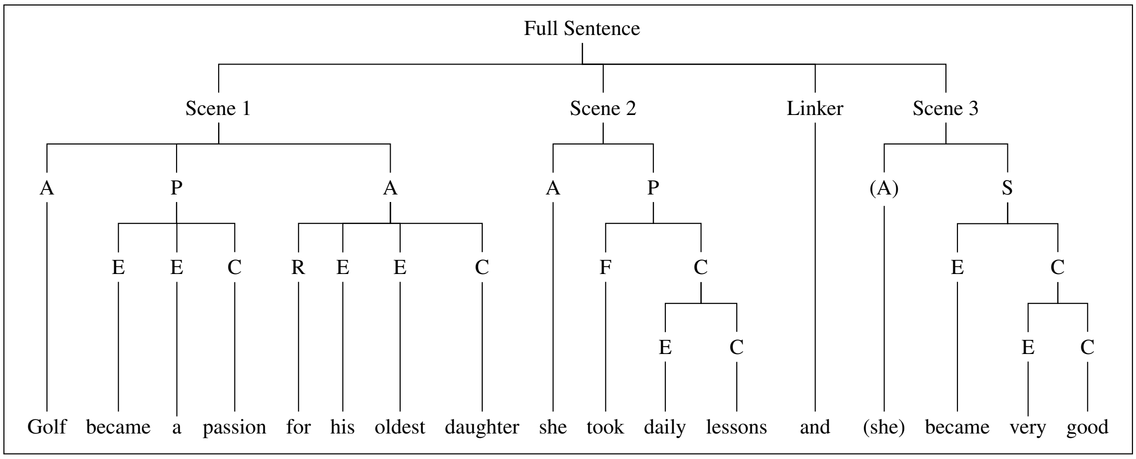


Figure 2.3: The HCOMET Converted Tree from the sentence annotated with UCCA in Table 2.3.

As mentioned above, for conversion into a constituency tree, the implicit participant *she* is made explicit.

[1] Functions include the *to* in infinitive verbs (e.g. *to*$_F$ *walk*) and auxiliary verbs (e.g. I am$_F$ running).

2.2.2 Alignment

After the application automatically transforms the UCCA representation of the reference and machine translations into their HCOMET counterparts, the translations must be aligned. Since the quality of translation is not binary, HCOMET allows node-alignments to be made completely or partially, meaning that certain nodes may be aligned even if they are not exact equivalents but do convey similar meanings. Whether two nodes are completely aligned or partially aligned may be subjective but human judgement as to the quality of similarity or semantic preservation is one of the advantages, not disadvantages, of human MT evaluation.

The HCOMET trees and their alignments offer much more information as to the quality of translation than pairwise ranking or even simple scoring of translation quality. From this information, it conceptually possible to derive various metrics of semantic preservation. The HCOMET project attempted to compile, from these data, elegant and cognitively-informed methods of semantic MT evaluation, three of which are defined in Chapter 4.

Chapter 3

Collection of Corpora

3.1 Source and Framework

3.1.1 Source Corpus

All discerning translation metrics must be able to compare various types of machine translation systems and work in different languages. Accordingly, the project used the output of two different machine translation systems selected from the 2013 WMT evaluation[1]: a phrase-based system (`uedin-wmt13`) and a rule-based system (`rbmt-3`) as well as the respective reference translations. These machine translation systems were used to translate from German to English as well from English to German. As such, the corpora allowed the testing HCOMET's language independence by running the analyses in both languages' and both systems' corpora.

Using these MT systems in English and German also allowed for direct comparison between HCOMET and HMEANT since Birch et al. (2013) employed these same machine translation systems and languages. Furthermore, the same number of English and German annotators were utilized.

3.1.2 Annotators, Training and Funding

In order to calculate IAA as well as compare HCOMET's usage in English to its usage in German, four English-language and two German-language annotators were utilized. Funding for German-language annotators was allocated from the European Union Seventh Framework Programme (FP7/2007-2015) under grant agreement 287658 (EU BRIDGE). The English-language annotators were researchers and students in the field

[1] `www.statmt.org/wmt13`

of Natural Language Processing. However, their backgrounds should not influence the reliability of the results, since the performance gap between annotators of different backgrounds “quickly vanished” after only five training passages in previous experiments (Abend and Rappoport, 2013).

All annotators completed a two-hour training in which UCCA and HCOMET concepts were explained and examples were discussed. Appendices A and B contain the information about HCOMET provided to annotators (for the UCCA annotation guidelines presented, contact Omri Abend). After the tutorial, German-language annotators were oriented to complete 38 hours of annotation and alignment while each English-language annotator was oriented to complete about 8 hours of annotation and alignment.

3.1.3 Structure of the Source Corpus

In order to provide annotators with the appropriate sentences to annotate and align, the source corpus was reorganized into three separate projects. Each project comprised a list of sentence pairs where the machine translation and its respective reference translation were grouped together. The list alternated between the two machine translation systems so that no matter how many sentence pairs were annotated an even number would come from each of the two MT systems, providing a better chance for statistically relevant results. Based on the number of English and German annotators, the projects were divided as shown in Table 3.1. The structure of the resulting annotated corpus is displayed in Table 3.2.

	English Project 1	English Project 2	German Project
Phrase Based MT & Reference Pairs	English Annotators 1 & 2	English Annotators 3 & 4	German Annotators 5 & 6
Rule Based MT & Reference Pairs			

Table 3.1: The Structure the Source Corpus Used

3.1.4 Annotation Time

Not all annotators worked at the same speed. In fact, while the fastest participant was able to annotate a translation pair in 7 minutes, the slowest annotator was 7 times

	Combined English	German Project
Phrase Based MT & Reference Pairs	31	11
Rule Based MT & Reference Pairs	33	10

Table 3.2: The Resulting Annotated Corpus: Number of translation pairs annotated by both participants per project, combining both English projects into one.

slower, averaging 49 minutes per translation pair. Such a wide range might signify that different approaches were taken and certain annotators were much more careful or unsure of how to annotate than others. All English-language annotators were faster than the German-language annotators, signifying that UCCA is perhaps more easily applied to English than to German. Additionally, the significantly worse machine translations in German may have caused the annotators to take more time to describe the nonsensical machine translations.

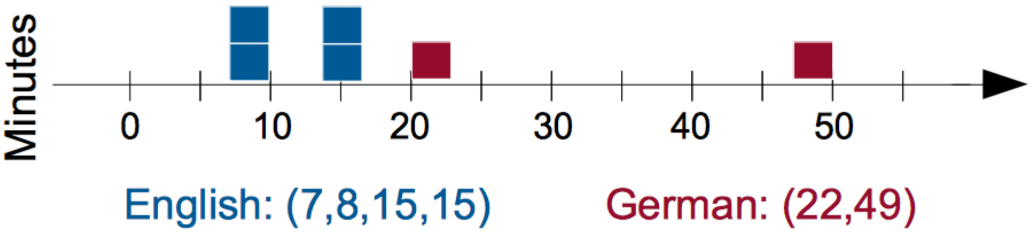


Figure 3.1: Number of Minutes per Average Annotation

3.2 Creation of HCOMET Annotation Tool

Developed incorporating a new version of the original UCCA annotation tool[2] (Abend and Rappoport, 2013), the HCOMET annotation web application was used for the parallel annotation and alignment of reference and translation pairs. The UCCA and HCOMET subapplications worked independently as indicated by the following steps:

[2] The original UCCA annotation tool can be found at `http://vm-05.cs.huji.ac.il/`

- Upon loggin in, a user is prompted to begin a new translation pair.

    1. Annotators use the UCCA Annotation Tool with the reference translation. The result is saved as an XML file and added to the database.
    2. Annotators use the UCCA Annotation Tool with the machine translation. The result is saved as an XML file and added to the database.
    3. The HCOMET Annotation tool loads the two XML files, converting the structures to HCOMET as described in Section 2.2, and allows users to align HCOMET nodes. The result is saved to a database.

- The user may log out and the data will remain in the database.

3.3 HCOMET Annotation Walkthrough

A visualization of the HCOMET annoation and alignment process is described in more detail as follows:

The annotation and alignment process, with an example originally given by Abend and Rappoport (2013) placed in the context of MT evaluation, can be summarized as follows: The user is given the reference translation to annotate. Using the UCCA annotation tool, the user deconstructs the sentence into scenes and linkers. The scenes are further annotated with the UCCA framework labels (Figure 3.2). Upon submission, the user will be prompted to complete the same steps for the machine translation (Figure 3.3). Once both sentences are fully annotated, the user is asked to employ the HCOMET application to align units from the reference to the MT output sentence. Each alignment must be marked as either complete or partial depending on how well the translation preserves the semantic integrity of the reference translation (Figure 3.4).

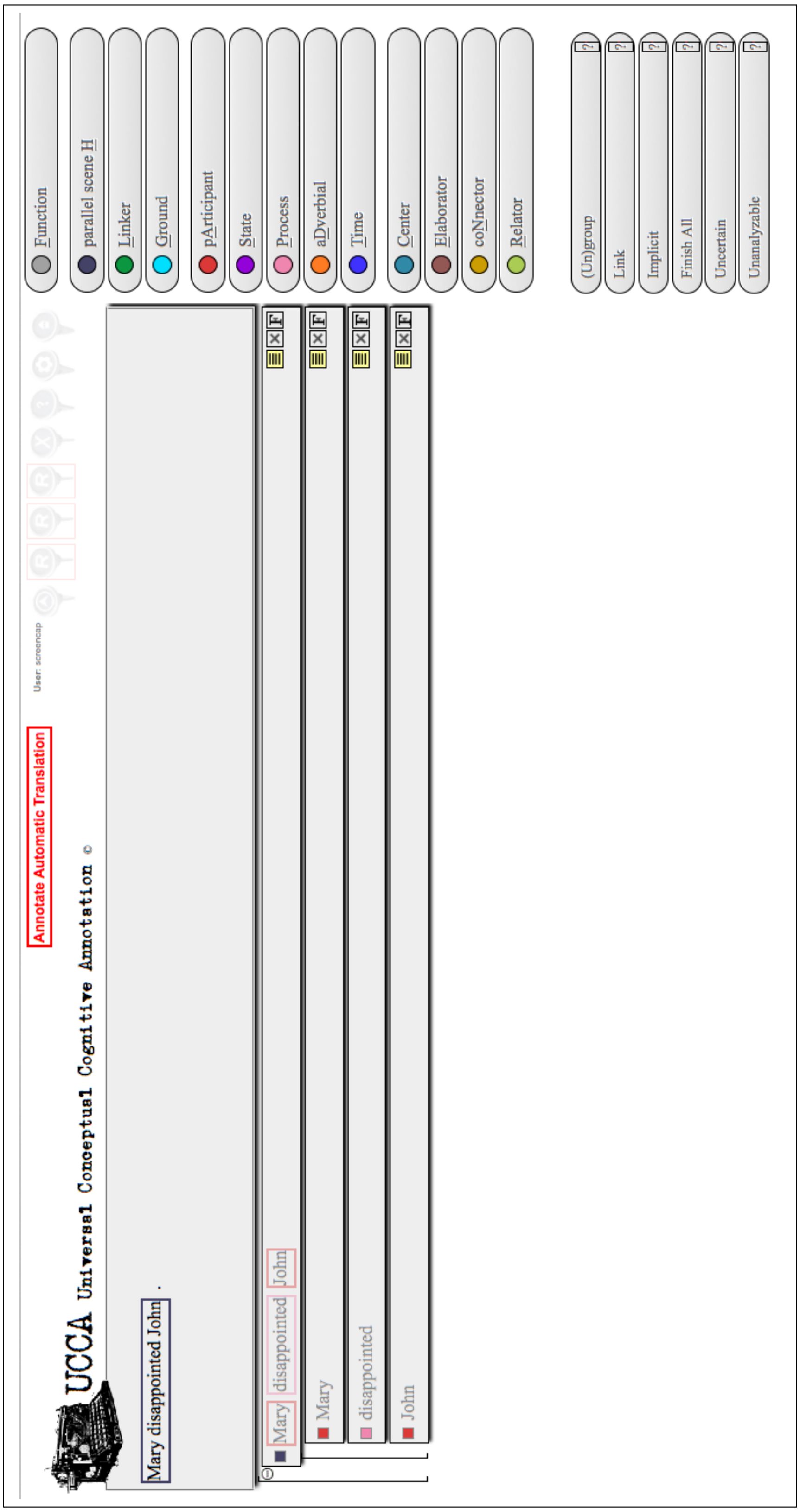


Figure 3.2: UCCA Annotation Tool: Annotation of Reference Translation

The annotator should become familiar with the reference translation to decrease the chance of using interfering but “non-contradictory (‘conforming’) analyses” of each version of translation (Abend and Rappoport, 2013). Starting with the reference translation, the annotator uses UCCA to divide the translated sentence into scenes and then into further elementary units.

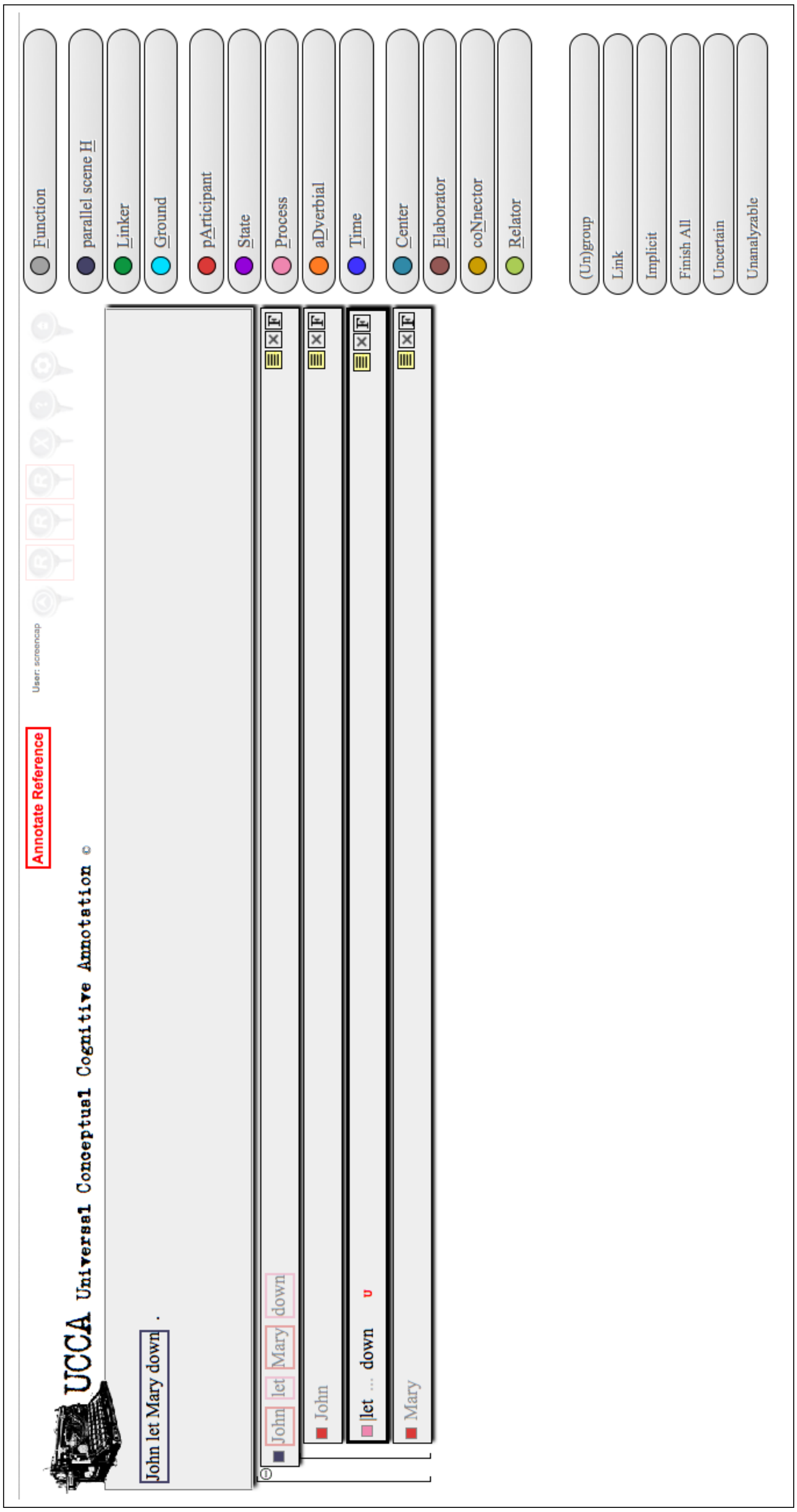
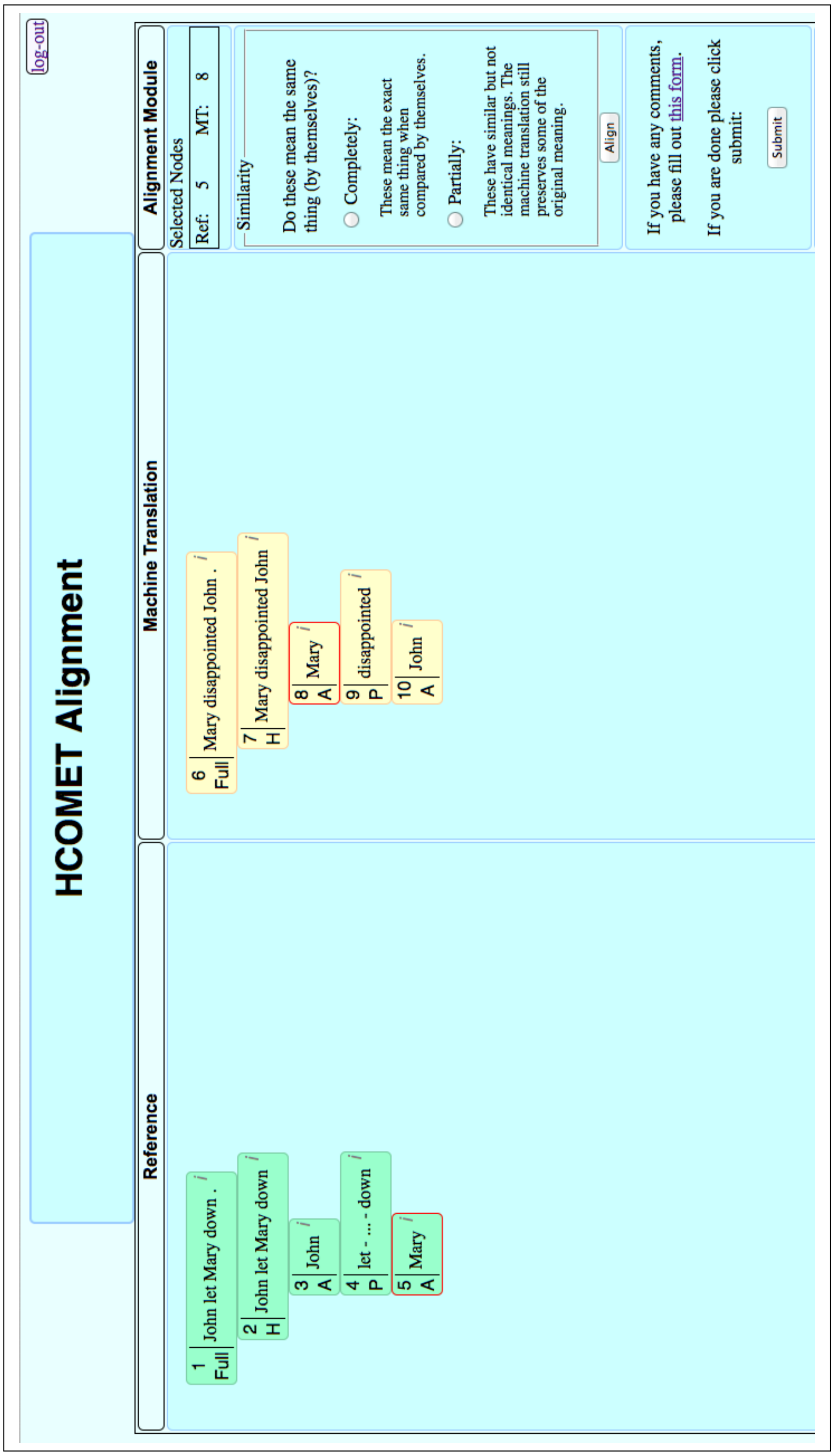


Figure 3.3: UCCA Annotation Tool: Annotation of Machine Translation

Continuing with the machine translation, the annotator again uses UCCA to divide the translated sentence into scenes and then into further elementary units.



Selected Nodes

Ref: 5 MT: 8

Similarity

Do these mean the same thing (by themselves)?

○ Completely:

These mean the exact same thing when compared by themselves.

○ Partially:

These have similar but not identical meanings. The machine translation still preserves some of the original meaning.

If you have any comments, please fill out this form.

If you are done please click submit:

Figure 3.4: HCOMET Alignment Tool: Alignment of Nodes

Annotators were oriented to begin aligning leaf nodes, moving up the tree until the Full sentence is aligned. Each alignment must be marked as Complete or Partial.

Chapter 4

Methodology of Analysis

After collecting the annotated corpora, the data were analyzed through several scoring and reliability metrics. The following sections outline the conceptual framework from which the analyses and their mathematical formulae were derived.

4.1 Calculating Translation Scores

4.1.1 Context of Numeric Scoring

A significant problem with the 2013 WMT's method of system ranking was that "the absolute value of the ranking or degree of difference [between translations was] not considered" (Bojar et al., 2013). Other values such as BLEU are difficult to interpret as semantically meaningful in spite of their ubiquity in MT research. This project, therefore, attempts to provide a numerical metric whose magnitude has linguistic and semantic relevance. Such a metric would also provide methods for absolute ranking and comparison. The HCOMET metric will attempt to follow the assertion that "[t]he closer a machine translation is to a professional human translation, the better it is" (Papineni et al., 2002), by calculating the semantic distance between the reference and machine translations.

Given that HCOMET annotations will be comprised of an UCCA-annotated reference translation tree, an UCCA-annotated machine translation tree, and the alignments between the two, a very large amount of data is available for analysis and compilation into a numeric value. The HCOMET metric attempts to elegantly compile this large amount of data into a single numeric value between 0 and 1 to represent the translation quality. A score of 0 would mean that no semantic content has been retained, while a score of 1 would mean the semantic content has been entirely retained.

Still a proof-of-concept, the HCOMET experiment will focus on the reliability and conceptual viability of the metric; as such, the mathematical formulae derived are conceptually elegant and do not attempt to assign weights to the features annotated. Many of the UCCA labels and arrangement patterns could carry weights specific to their relevance in semantic preservation. However, this kind of optimization will be left for future research.

A total of three metrics were derived to encapsulate different aspects of translation quality. The HCOMET metric measures overall translation quality; the LEAF score measures how well individual words or primary semantic units are translated; the SCENE score calculates the preservation of whole scenes in a translation.

4.1.2 Compositionality Semantics and Semantic Equivalence

As mentioned in the introduction, HCOMET attempts to lower the cognitive complexity of comparing translations by breaking them down into smaller units of meaning. However, in order to calculate an overall numeric score for translation quality we must recombine all of these units into a single numeric value. In order to so do an understanding of semantic equivalence and compositionality semantics is required.

Simply stated, “[f]or two sentences α and β, if [in any possible situation] α is true and β is false, α and β must have different meanings” (Cresswell, 1982). However elegant this is, it only provides a binary distinction between equivalence and inequality. This dichotomy does not provide a sufficient description for MT systems. Consequently, we apply this idea to each of the smaller units provided by the HCOMET annotations and thus describe which parts of the sentence are translated well and which parts are not. However, two limitations still persist: Even at the individual lexeme-level, translation quality is not binary but may fall within a range, and the overall quality of translation cannot be derived from merely aligning the leaf nodes since “ the meaning of the whole is a greater than the meaning of the parts” (Lakoff, 1972).

The limitation of binary equivalence is addressed by HCOMET’s partial and complete alignments. Human annotators may leave translated nodes unaligned, meaning that they do not represent any intelligible aspect of the reference translation; they may completely align translated nodes, if they are a perfect translation; or they may partially align nodes, if they capture some, but not all, of the semantic content of their respective reference nodes.

The other limitation, compositionality, has been previously addressed in NLP using the idea that "composition of simple elements must allow the construction of novel meanings which go beyond those of the individual elements" (Pinker, 1994) through methods ranging from vector multiplication to lambda calculus (Mitchell and Lapata, 2009). More linguistically motivated studies have attempted to use the way units are "syntactically combined" as a method to capture the novel meanings created (Partee, 1995); yet, research such as Ge and Mooney (2009) has shown that internal syntactic structures do not always represent the internal semantic structures. Humans, however, can inherently derive these novel meanings. The H (human) in HCOMET is, therefore, an integral part in being able to compare semantic compositions at all levels of the semantic tree annotated through UCCA.

Thus, HCOMET annotations have an advantage both in terms of equivalence and compositionality since UCCA trees are inherently semantic and human annotators are much more adequate at defining equivalence. Conceptually, HCOMET attempts to use human definitions of equivalence and a cognitively-informed method of semantic compositionality to create a function which follows Frege's Principle of compositionality: "The meaning of whole is a function of meaning of its parts" (Partee, 1995).

4.1.3 Devising Mathematical Formulae

The following scoring guidelines and formulae define the function for individual sentences (a single reference and machine translation pair). This was done so that the conceptual framework could be more easily seen through its derived mathematics. However, the scores reported were all calculated at the corpus (individual machine translation system) level. Such corpus-level scores are not simple averages of all sentences in the corpus since shorter sentences and longer sentences should not be given the same weight. In order to properly convert between sentence level and corpus level, the precision, recall, and F1 combination formulae are calculated for the whole corpus instead of each sentence.

BLEU

Although BLEU is not directly comparable to HCOMET since it is an automatic metric, its ubiquity in machine translation research provides a useful benchmark with which HCOMET and other scores can be compared. Corpus-level, 4-gram BLEU[1]

[1] Sennrich (2014)'s implementation was used.

scores were computed for each MT system's output given each sentence's single reference translation. By comparing BLEU with HCOMET, it is possible to determine n-gram-based translation metrics and semantic-based translation metrics differ.

HCOMET Score

In order to provide an overall numeric value of translation quality, the HCOMET score uses a simple recursive definition in order to compile the annotation alignments. The root node used at the top level recursion is the first node in the HCOMET tree with more than one child. For example, if a $Full$ node only father's a single H node, the H node is chosen as the root since the unary child is an artifact of the HCOMET application's internal architecture.

Root Nodes:

$$\mathrm{MT_r} = \text{The Root (Top) Node of the Machine Translation}$$

$$\mathrm{REF_r} = \text{The Root (Top) Node of the Reference Translation}$$

Size:

$$n(node) = 1 + \sum n(child)$$

Score:

$$s(node) = \begin{cases} n(node) & \text{if completely aligned} \\ 0.5 + \sum s(child) & \text{if partially aligned} \\ 0 & \text{if not aligned} \end{cases}$$

% Correct:

$$c(node) = \frac{s(node)}{n(node)}$$

Precision & Recall:

$$P = c(\mathrm{MT_r})$$

$$R = c(\mathrm{REF_r})$$

HCOMET Score:

$$HCOMET = \frac{2 * P * R}{P + R}$$

LEAF Score

LEAF (Leaf Equivalence Assessment Function) attempts to determine how well a translation chooses individual leaf translations. This scoring guideline does not take into account the composition of the translation but merely individual leaf translations. By comparing LEAF scores to HCOMET scores it is possible to determine whether a machine translation system is performing poorly because of inadequate lexical translations or because of improper semantic compositionality[2].

Leaf Nodes:

$$\text{MT}_{\text{leaves}} = \text{The Leaf Nodes of the Machine Translation}$$
$$\text{REF}_{\text{leaves}} = \text{The Leaf Nodes of the Reference Translation}$$

Node Score:

$$s(leaf) = \begin{cases} 1 & \text{if completely aligned} \\ 0.5 & \text{if partially aligned} \\ 0 & \text{if not aligned} \end{cases}$$

% Correct:

$$c(\text{leaves}) = \frac{\sum\limits_{\text{leaves}} s(\text{leaf})}{\text{number of leaves}}$$

Precision & Recall:

$$P = c(\text{MT}_{\text{leaves}})$$
$$R = c(\text{REF}_{\text{leaves}})$$

LEAF Score:

$$LEAF = \frac{2 * P * R}{P + R}$$

Scene Score

SCENE (Scene Complete Equivalence Numeric Evaluation) attempts to provide a shallow, top-down assessment of the translation quality. By averaging the alignment

[2] The LEAF metric is reminiscent of METEOR in terms of unigram alignment (Banerjee and Lavie, 2005).

values of solely the scene nodes[3], SCENE scores capture whether machine translation systems are able to retain the internal structure of individual scenes without taking into account how the scenes are arranged in the full translation. By ignoring discourse markers (such as *L* and *G* nodes), the SCENE score represents whether MT systems can preserve individual scenes.

Scene Nodes:

$$\text{MT}_{\text{scenes}} = \text{The Scene Nodes of the Machine Translation}$$

$$\text{REF}_{\text{scenes}} = \text{The Scene Nodes of the Reference Translation}$$

Node Score:

$$s(leaf) = \begin{cases} 1 & \text{if completely aligned} \\ 0.5 & \text{if partially aligned} \\ 0 & \text{if not aligned} \end{cases}$$

% Correct:

$$c(\text{scenes}) = \frac{\sum_{\text{scenes}} s(\text{scene})}{\text{number of scenes}}$$

Precision & Recall:

$$P = c(\text{MT}_{\text{scenes}})$$

$$R = c(\text{REF}_{\text{scenes}})$$

SCENE Score:

$$SCENE = \frac{2 * P * R}{P + R}$$

4.2 Calculating HCOMET Inter-Annotator Agreement

4.2.1 IAA and Reliability of Metrics

Although UCCA required post-processing in order to calculate its IAA, since there is “no standard evaluation metric for comparing two grammatical annotations in the form of labeled [Directed Acyclic Graphs]” (Abend and Rappoport, 2013), HCOMET already processes UCCA into tree structures. Thus, the HCOMET constituency trees can employ previously established methods of calculating IAA.

[3] “A Scene can describe some movement or action, or a temporally persistent state” (Abend and Rappoport, 2013). In effect, scene nodes are all *H* nodes and the *A* or *E* nodes that immediately father *S* or *P* nodes.

The mathematical means to calculate IAA will be adapted from Birch et al. (2013) and Lo and Wu (2011) so that the numbers can be comparable and the robustness of HCOMET as a metric of semantic preservation can be directly contrasted with other schema. Both of these studies utilized a method of computing IAA by adapting labeled precision and recall of syntactic parsing. One annotation is treated as the the gold standard from which the second annotation differs. By calculating the F1 score of the second annotation, it is possible to calculate how similarly the two trees are constructed.

IAA:

$$\text{Mutual Nodes} = \text{Nodes that cover the same text}^{4}$$

$$P = \frac{\text{\# mutually annotated nodes}}{\text{\# nodes in annotation of Annotator 1}}$$

$$R = \frac{\text{\# mutually annotated nodes}}{\text{\# nodes in annotation of Annotator 2}}$$

$$IAA = \frac{2 * P * R}{P + R}$$

[4] Mutual Nodes are calculated ignoring child Fs, Rs, and remote text.

Chapter 5

Results and Discussion

As previously mentioned, UCCA, HCOMET's annotation scheme, is based on principles of cognitive linguistics, abstracting semantic values away from syntax-based frame semantics. Specifically, UCCA's framework builds upon Langacker's basic tenets of cognitive grammar, that "lexicon and grammar form a gradation consisting solely in assemblies of symbolic structures" and that syntax is not autonomous or distinct from either lexicon or semantics (Langacker, 2008). Therefore, UCCA's scenes and labels are not constrained by syntactic variation or restriction. Although these benefits seem to be directly applicable to MT evaluation, the linguistic properties of UCCA have not yet been tested in the domain of machine translation. Thus, for a comprehensive analysis of the linguistics properties of HCOMET, this chapter is divided into three sections:

1. HCOMET's theoretical ability to overcome hindrances[1] commonly faced by frame semantics and SRL-based evaluation metrics such as HMEANT (Birch et al., 2013) is demonstrated.

2. These theoretical abilities are experimentally applied and reported through HCOMET scores, LEAF scores, and SCENE scores.

3. HCOMET's reliability is analyzed through its Inter Annotator Agreement.

[1] Hindrances such as reliance on verb-heads and unclear hierarchy of frames.

5.1 Linguistic Properties and Potential of HCOMET

5.1.1 Linguistic Comparison to Frame Semantics and HMEANT

Frame semantics attempts to use SRLs to calculate how well a translation retains the semantic content of its original sentence. However, the HMEANT experiment described several linguistic phenomena that did not fit into the frame-semantic infrastructure (Birch et al., 2013). Below is a list of these phenomena and how HCOMET (through UCCA) is able to capture the semantic structure of setences more accurately.

Phrasal Verb Heads

In English, phrasal verbs such as *wake up* are frequently split in sentences such as *Please, wake him up!*. HMEANT, however, was not able to annotate single but discontinuous units of meaning. This is a severe problem in English, but also occurs in other languages such as German through separable verbs. For example, *ankommen* is split in the sentence *Sie kommt sofort an*. UCCA's ability to annotate discontinuous nodes and the ability to mark multi-word nodes as semantically unanalyzable allow annotators to bypass word-order and even word parsing in order to annotate the appropriate semantic content of a structure. An example can be seen in node 4 of Figure 3.4.

Non Verbal Heads and Light Verb Constructions

In many languages, including English, predicates can be created using light verb constructions, in which the syntactic verb carries very little semantic value. While the syntactic verb of the sentence *Mary napped all afternoon* carries the appropriate predicate meaning, the syntactic verb *took* of an equivalent sentence, *Mary took a nap all afternoon*, does not. HMEANT would require the frame of the latter sentence to be centered around *took* which would make it not align with a frame centered on *napped*. The fact that UCCA makes both the verb *napped* and noun *nap* the processes[2] of their respective scenes allows for more accurate semantic annotations.

Copular Verbs

Forcing a sentence such as *The boy is young* to have an agent and a patient, which is the case in HMEANT, creates a structure that does not represent the true semantic value

[2] Processes (P) or Centers (C) of Processes

of, specifically, the so-called patient. By having a predicate to represent stative scenes, UCCA accounts for these sentences and allows annotators to demarcate *young* as the State (S)[3]. HCOMET's removal of Function (F) nodes further aligns the semantics of the sentence to the structure of the tree shown in Figure 5.1 since copulas do not contain semantic content.

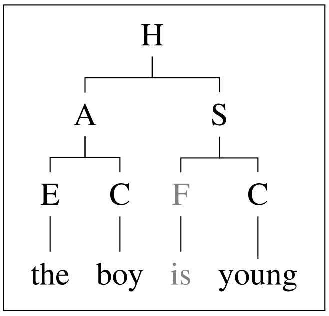


Figure 5.1: Example of State (S) in UCCA/HCOMET

Preposition Phrase Attachment

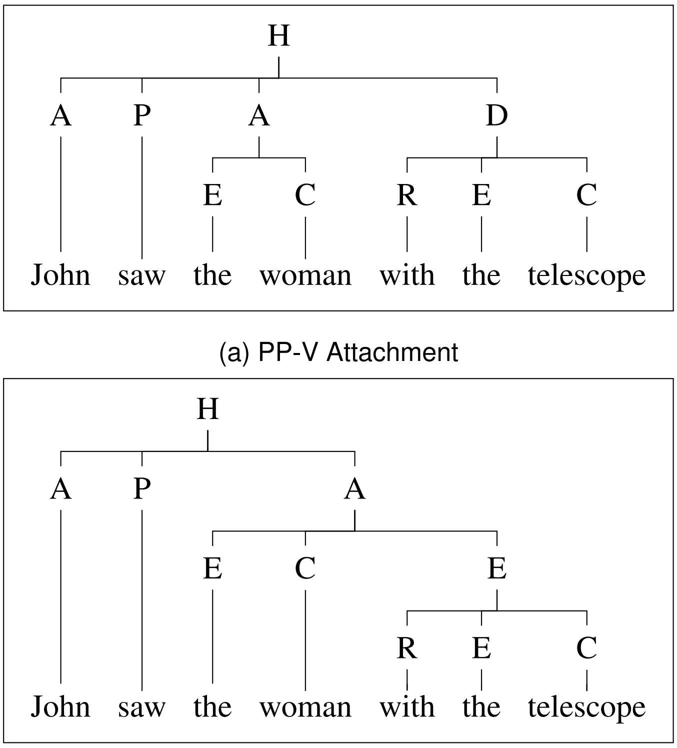
While the prepositional phrase (PP) can elaborate the semantics of both nouns and verbs, all roles in HMEANT were linked to the predicate center of the frame. Therefore, "HMEANT has no way of capturing this [difference]" (Birch et al., 2013), meaning that the attachment of the PP was ignored when scoring. Much like a syntactic tree, HCOMET's semantic tree does provide the means to annotate to what a PP refers. Figure 5.2 demonstrates, respectively, a PP-verb attachment and a PP-noun attachment on an ambiguous sentence.

Hierarchy of Frames

Although HMEANT can define when frames perform a role within another frame, the scoring guidelines do not account for this, given that the precision and recall are simply averages of the frames in each sentence (Birch et al., 2013). HCOMET uses its tree structure in order to create a more elegant, recursive definition for its main score. Therefore, HCOMET does not need to average the scenes of an annotation since they are automatically weighed through its recursive definition.

In the HMEANT experiment it was "not clear whether errors at the lowest level" such as within reported speech "should be marked wrong just at that point, or whether

[3] State (S) or Center (S) of a States.


H
A P A D
E C R E C
John saw the woman with the telescope
H
A P A
E C E
R E C
John saw the woman with the telescope

(a) PP-V Attachment

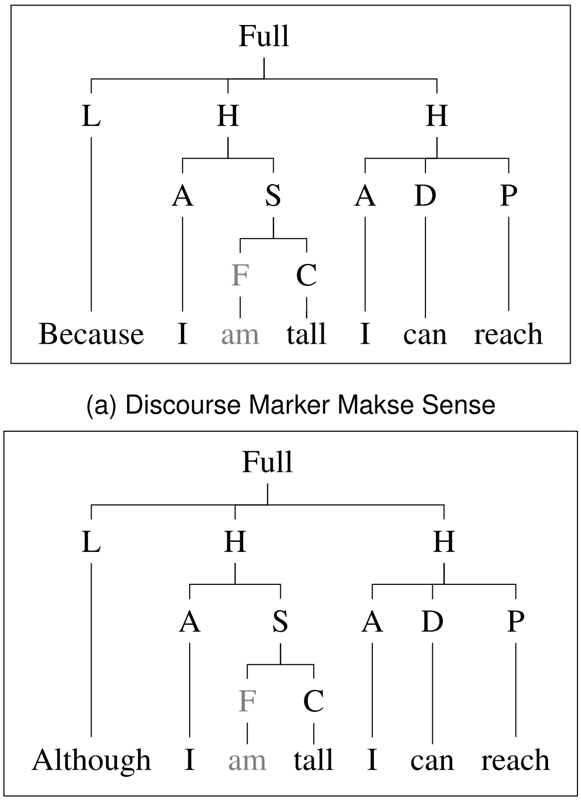
(b) PP-N Attachment

Figure 5.2: Example of PP Attachment in UCCA/HCOMET

they should be marked wrong all the way up the semantic tree" (Birch et al., 2013). Humans do not want to mark a frame as completely correct if one of its internal components is not completely correct; yet, the internal error has already been calculated in the subordinate scene. In the HCOMET experiment, the human aspect of the metric was embraced in an attempt to increase HCOMET's ease of use. Annotators were clearly instructed to discount errors all the way up the semantic tree.

Discourse Markers

HMEANT does not annotate discourse markers since they lie outside predicate frames, but "[t]hese are important for capturing the relationships between frames and should be labelled" (Birch et al., 2013). Since once of the fundamental units of UCCA is the Parallel Scene (H), UCCA does capture the relationship between them through the use of Linkers (L) and Grounds (G) as shown in Figure 5.3. The score's indiscriminate recursive definition does account for the alignment between the discourse markers in each sentence.



(a) Discourse Marker Makse Sense

(b) Discourse Marker Does Not Make Sense

Figure 5.3: Example of Discourse Marker Annotation UCCA/HCOMET

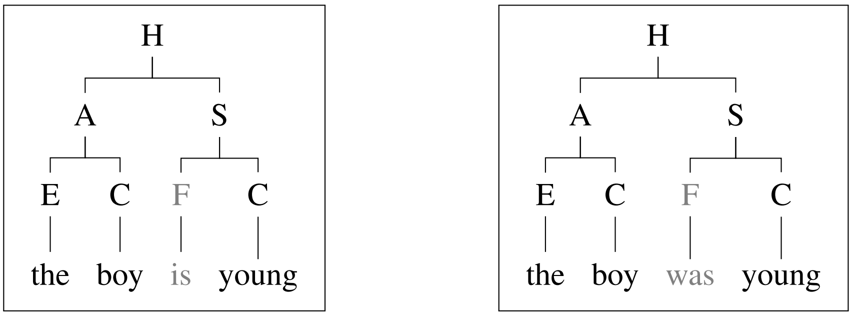
5.1.2 Other Linguistic Repercussion of HCOMET Properties

Tense and Redundancy

Since HCOMET, before the alignment step, deletes the UCCA Function (F) nodes in order to preserve only the semantic bearing nodes, certain oddities arise when a node contains two children which are a Function (F) and a Center (C). Although it might seem that the deletion of the F node would create a useless unary parent whose child will always bear the same alignment, this is not the case.

UCCA does not inherently account for tense at this point. However, HCOMET, in order to provide a true metric of semantic translation quality, must. This is where having an extra node derived from the existence of the Function (F) proves useful. UCCA would annotate the sentences in Figure 5.4 equivalently but they are inherently different because of their tenses.

Although the Centers (C) would be aligned completely, the States (S) would not,

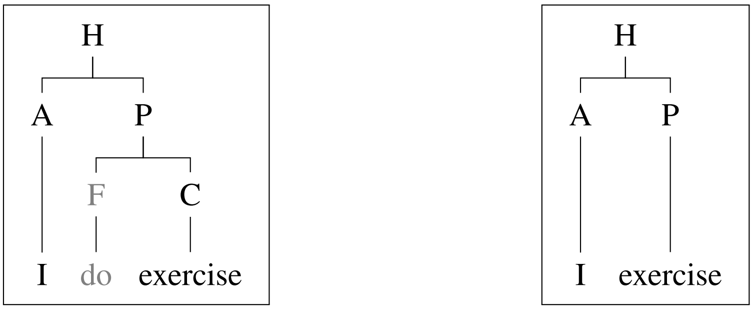


(a) Present Tense

(b) Past Tense

Figure 5.4: Example of UCCA's Function (F) Nodes Allowing HCOMET to Account for Tense.

precisely because of the extra level in the hierarchy derived from the original F node. This benefit, however, does not always exist. In certain cases such as in Figure 5.5 this extra layer might introduce redundancy into the HCOMET calculations. Since the Processes (P) would be completely aligned, the sentence with the Center (C) would give the aligned process the extra weight of the internal Center (C).



(a) Auxiliary Verb

(b) No Auxiliary Verb

Figure 5.5: Example of UCCA's Function (F) Nodes Introducing Redundancy to HCOMET.

One-to-One Alignment

Because HCOMET currently requires a one-to-one alignment between reference and MT nodes, some machine translations may be over-penalized if they inaccurately unite or separate reference nodes. If a reference scene is divided into two partially accurate machine-translated scenes, one of the machine-translated scenes will, necessarily, remain unaligned, even though it should be partially aligned.

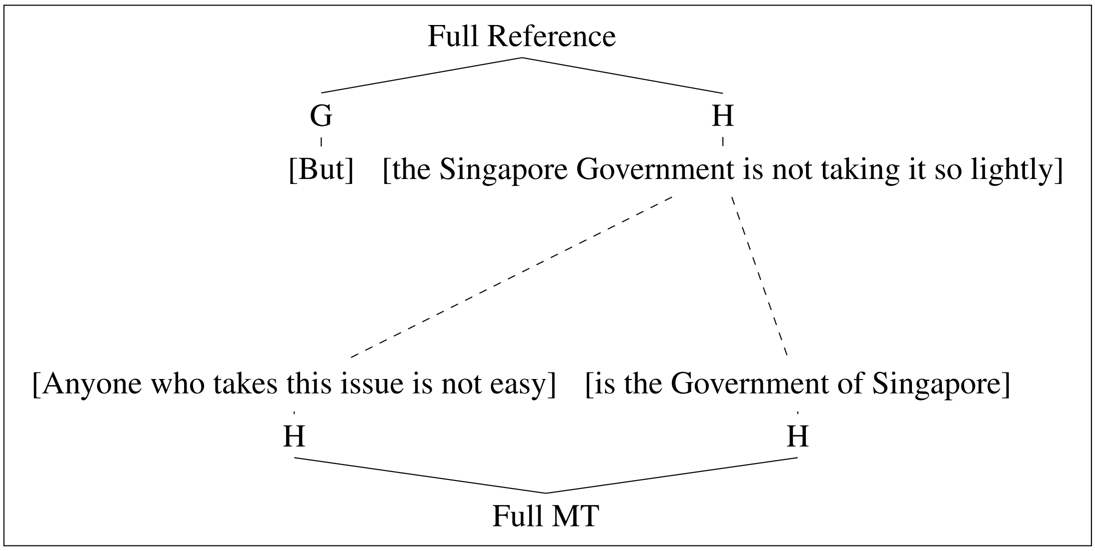

Full Reference
G
H
[But]
[the Singapore Government is not taking it so lightly]
[Anyone who takes this issue is not easy]
[is the Government of Singapore]
H
H
Full MT

Figure 5.6: Unequal Number of Translated Nodes

As Figure 5.6 demonstrates, even though both machine-translated scenes (H) carry some of the original meaning of the single reference scene, only one may be currently aligned giving the other scene a score of 0 as if it were untranslated.

5.2 Scores and Other Quantitative Results

Given the theoretical benefits of using HCOMET as a machine translation evaluation metric, 4 English-language annotators and 2 German-language annotators were asked to participate in an experiment whose results are presented in this section.

5.2.1 HCOMET, LEAF, & SCENE Scores

German to English Translation Results

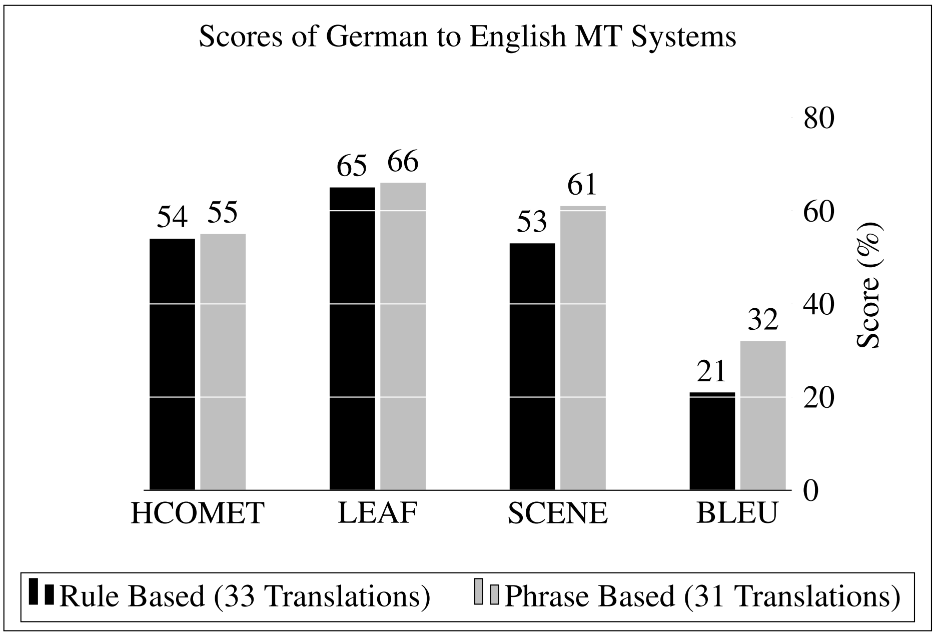


Figure 5.7: Calculated Scores for German to English MT Output.

Score Summaries:

BLEU

While BLEU scores are not directly comparable to the HCOMET system scores, it was calculated to see how a commonly used automatic MT metric might differ from the scores given by HCOMET. As might be expected, BLEU scored the phrase-based MT system higher than the rule-based system. What was noteworthy was the large 11-point difference.

HCOMET

HCOMET, on the other hand, deemed that the two systems were much closer in quality than BLEU, giving the phrase-based system a negligibly higher 1 percentage point. The lack of a substantial difference between the two systems might be significant since it may indicate that the rule-base system is unfairly penalized by BLEU because of its higher reordering, even if it retains the same amount of semantic content.

LEAF

Unexpectedly, the phrase-base system did not seem to do better in translating the individual leaf nodes. Having longer phrases and therefore more context from which lexical ambiguity could be solved should give the phrase-based system an advantage, but the LEAF Score demonstrates that this is not necessarily the case in the corpus tested.

SCENE

The only experimental score in which the two systems differ noticeably is the SCENE score. Broadly measuring how well systems translate full scenes, this score perhaps demonstrates that the phrase-based system does use its context-sensitivity to translate whole scenes in a more accurate fashion.

System Summary:

Phrase-Based MT System

The phrase-based MT system received slightly higher scores for the overall HCOMET score and for the LEAF Score; however, these differences were probably not statistically significant given their mere 1% magnitude. This system, might have an advantage in that it translates whole scenes more accurately[4]. This difference, however, does not seem to affect the overall scores perhaps resulting from a lower ability to arrange the translated scenes.

Rule-Based MT System

Although the rule-based MT system rated lower for all the scores, the differences were not as large as the BLEU scores would originally indicate. Such a discrepancy might signify that the rule-based system achieves as much translation as the phrase-based system, but in a manner that is penalized by the structure of BLEU.

[4] See Figure 5.9

English to German Translation Results

As expected, all scores when translating away from English were lower. However, the HCOMET, LEAF and SCENE scores all ranked the rule-based system as more semantically sound, while BLEU deemed the phrase-based system better.

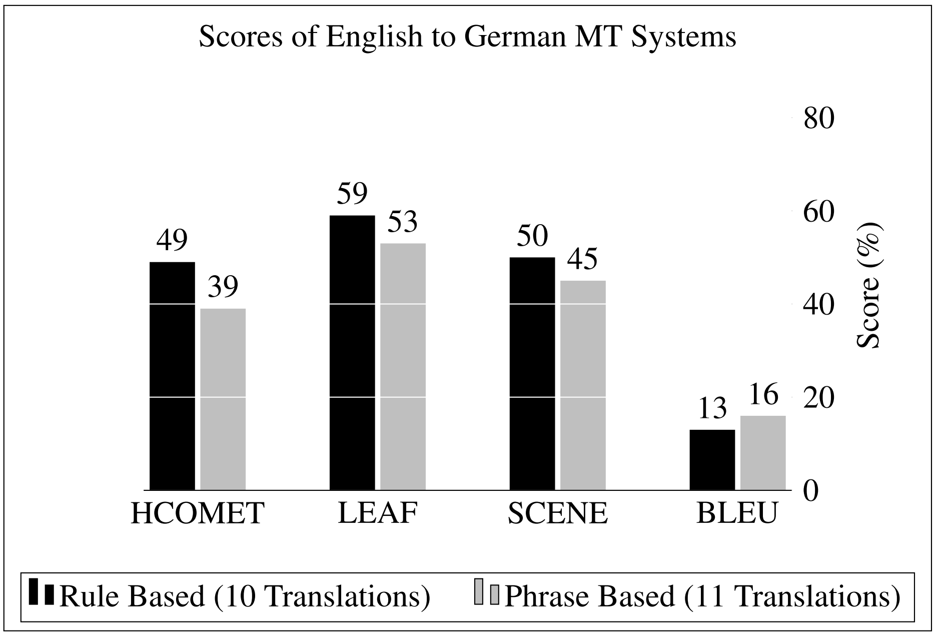


Figure 5.8: Calculated Scores for English to German MT Output.

Score Summaries:

BLEU

BLEU, again, gave the phrase-based MT system a higher score, though by a much smaller margin. In this case, however, BLEU stood out in ranking the phrase-based MT system as more accurate. All three other scoring guidelines determined that the rule-based system provided more accurate translations. Perhaps BLEU was incorrectly influenced by correctly translated n-grams while human annotators did not understand their misplacement in German.

HCOMET

A full 10% difference was found between the HCOMET scores of the rule-based and phrase-based MT systems. Although the difference between the LEAF scores and the SCENE scores were not as large, the rule-based MT system must have preserved semantically required reordering of the leaves and scenes, whereas the phrase-based system did not.

LEAF

In spite of the reordering issues found in phrase-based translation systems, one would expect that the phrase-based translation's individual words would be translated more accurately; in the corpus, however, this was not the case.

SCENE

The SCENE scores presented the smallest difference, perhaps signifying that although the phrase-based system might preserve individual scenes almost as well as the rule-based system, something else such as global reordering might be affecting the overall translation quality.

System Summary:

Phrase-Based MT System

While the phrase-based MT system received a slightly higher BLEU score, it seems that it did not preserve the semantics of the reference translation as well as the rule-based system. This can be seen in all three HCOMET, LEAF, and SCENE scores. The higher BLEU score must be an artifact from misplaced but still accurate n-gram translations.

Rule-Based MT System

The LEAF and SCENE scores of the rule-based MT system were higher than those of the phrase-based MT system. However, these differences do not account for the entire 10% difference in the overall HCOMET scores. The rules found in the English to German rule-based MT system must preserve semantically-required reorderings not found in the phrase-based MT system.

5.2.2 Further Results

Global scores are useful in determining the overall semantic quality of a machine translation system. However, global scores might reduce complexities in the data which may provide useful information for further developments. Given the large amount of information gathered through the HCOMET annotations and alignments, we are not only able to calculate single value percentage scores, but we are also able to calculate the specific semantic effects of specific translation systems.

For example, although we weigh translation quality in terms of full and partial alignments, it is useful to understand how many sentences are translated perfectly and how many sentences are completely unintelligible. There is a significant difference between a system that translates all sentences imperfectly and a system than translates half the sentences perfectly and the other half unintelligibly. Thus, the exact proportion[5] of translation qualities was calculated and is displayed in Figures 5.9 and 5.10.

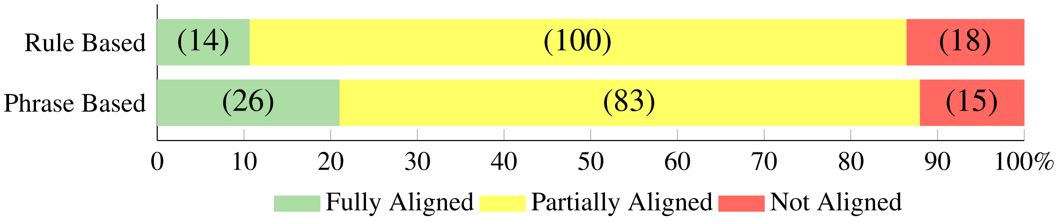


Figure 5.9: English Alignment of Full Sentences: Proportion and (Count)

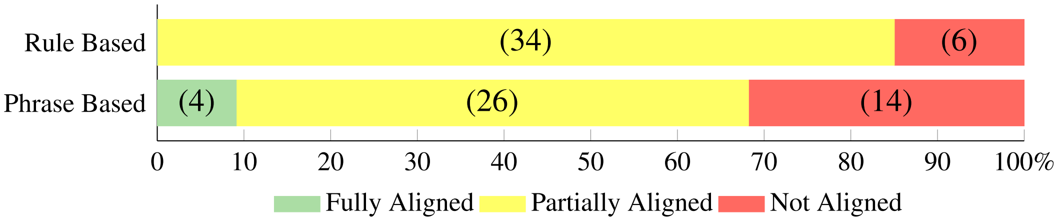


Figure 5.10: German Alignment of Full Sentences: Proportion and (Count)

Furthermore, it may be of interest to see how individual systems translate specific types of semantic constructions. If a system is built to emphasize correctly translated actions, perhaps the proportion of correctly translated Processes (P) would be of use. Because all of this information is available after HCOMET annotation and alignment, we have the ability to produce the proportions found in Figures 5.11 and 5.12.

Appendices C and D contain the proportions for all UCCA types for the German to English systems and the English to German systems, respectively.

[5] The counts depicted add up to four times the number of translation pairs because each project has two annotators and each annotator has a reference and a machine translation to annotate.

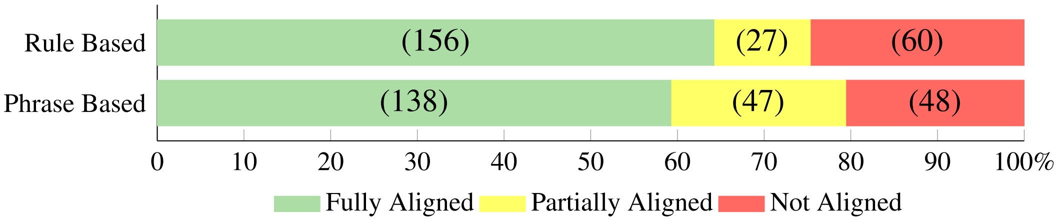


Figure 5.11: English Alignment of Processes: Proportion and (Count)

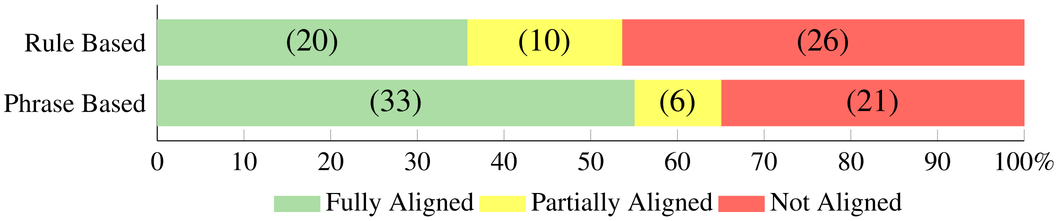


Figure 5.12: German Alignment of Processes: Proportion and (Count)

5.3 Reliability of the Metrics

However interesting the score results may be, they do not mean anything unless the method of scoring is reliable. This is why along with the scores, the inter-annotator agreement was calculated and is presented in this section.

5.3.1 Node Identification & IAA

Asking participants to construct a semantic tree, given the UCCA guidelines, was the first step in order to calculate HCOMET scores. This step, however, was not trivial since semantic trees are not always equivalent to the syntactic trees taught in school. Therefore, subjects may have been confused with the concept of a semantic tree. Table 5.1, nonetheless, demonstrates that this was not the case. With IAA F1s ranging from 79% to 83%, we can confidently say that annotators with only two hours of training were able to construct semantic trees. The IAA F1s in the HMEANT experiment ranged from 73% to 76% with only the German reference translation IAA F1 surpassing UCCA, though only by 1.3% to achieve 84.6%.

Output Language	Reference Translation	Machine Translation	Overall
English	0.83	0.81	0.82
German	0.83	0.79	0.81

Table 5.1: Unlabeled Inter Annotator Agreement (Node Identification) F1. The darker cells indicate a higher IAA.

5.3.2 Node Labeling & IAA

Knowing that participants can construct unlabeled semantic trees is not all that the HCOMET task required. Annotators must be able to use the UCCA types in order to provide a more detailed semantic representation of the translations. Therefore, the labeled IAA of the UCCA trees is also important. Table 5.2 demonstrates that although the IAA is not as high as HMEANT's role-classification, annotators still agreed in the majority of cases and, perhaps, with more training this IAA could increase as well.

Output Language	Reference Translation	Machine Translation	Overall
English	0.68	0.67	0.67
German	0.62	0.59	0.60

Table 5.2: Overall Inter Annotator Agreement F1. The darker cells indicate a higher IAA.

Inter Annotator Agreement per UCCA Type

Inter-annotator agreement was relatively stable between the reference and machine translations, meaning that the quality of translation did not affect the IAA. However, not all UCCA types were annotated as successfully. Table 5.3[6] demonstrates the high variation in IAA by language and UCCA type.

Perhaps caused by different levels of emphasis during training, certain UCCA types such as States (S) and Grounds (G) were not as reliably annotated as Elaborators (E) and Participants (A). These differences could be taken into account in future UCCA

[6] One of the German Annotators annotated Linkers (L) and Grounds (G) as Functions (F). Therefore, the IAA of these was 0.

Type		English	German
H	Parallel Scene	0.53	0.62
P	Process	0.60	0.50
S	State	0.28	0.14
A	Participant	0.67	0.71
D	Adverbial	0.49	0.33
T	Time	0.53	0.69
G	Ground	0.19	0.00
C	Center	0.71	0.70
E	Elaborator	0.73	0.61
N	Connector	0.69	0.74
R	Relator	0.77	0.36
L	Linker	0.61	0.0
Meta-Analysis			
Scene Units		0.47	0.59

Table 5.3: Inter Annotator Agreement of Annotations by Type F1. The darker cells indicate a higher IAA.

tutorials. For example, Such types as States and Grounds could be further emphasized to increase reliable usage and IAA.

5.3.3 Learning Curve and IAA

Given that annotators were only offered a two-hour tutorial on UCCA and HCOMET before starting their annotations, it was expected that their annotations would not perfectly follow the tutorial guidelines. In fact, once annotating, many participants were frustrated with not understanding how to annotate certain types of constructions. However, these frustrations lessened as each annotator gained experience. Since the order of sentences annotated by each participant was static, the indexed IAA reflected the annotators' increased experience and facility with UCCA and HCOMET.

Although the R^2 values of the linear regressions shown in Figures 5.13 are low, they suggest the possibility of a much higher IAA through increased training and experience. The numbers described in this section may be used as a benchmark for future IAA, given the short tutorial offered and annotators' lack of UCCA and HCOMET experience.

The average difference in HCOMET scores between the two annotators was relatively low, especially for the translations into English as shown in the histogram in Figure 5.14. Such a pattern may signify that not only do the structure of the annotated UCCA trees align, but the calculated HCOMET score aligns as well. However, the scores for the translation into German exhibit a much smaller degree of alignment, meaning that perhaps a higher structural UCCA IAA is required before HCOMET scores align at sufficiently high levels.

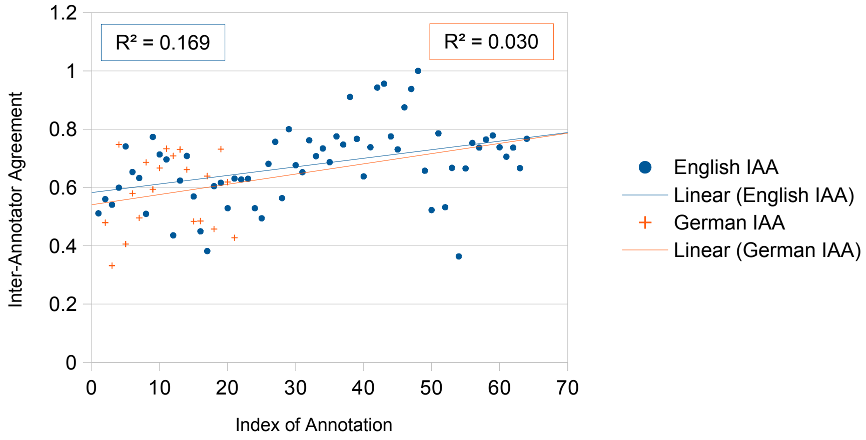


Figure 5.13: Learning Curve of HCOMET through Inter Annotator Agreement

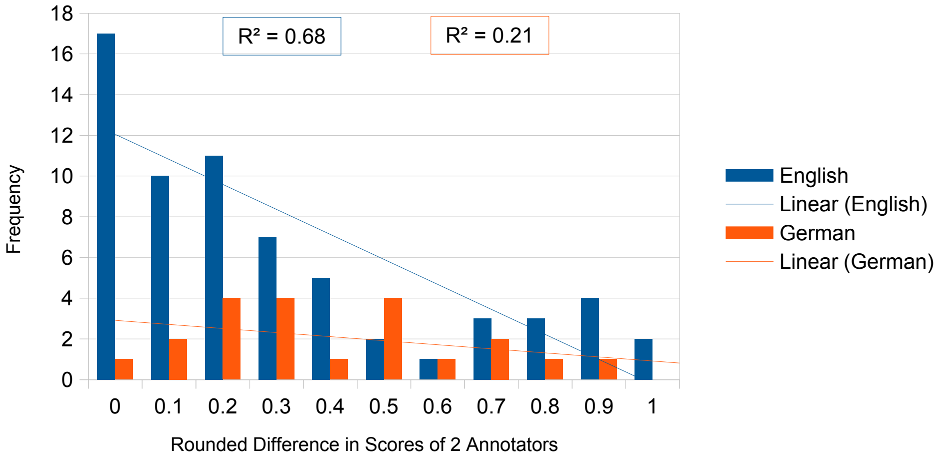


Figure 5.14: Difference Between HCOMET Scores

Chapter 6

Conclusion

Using an "intuitively natural, psychologically plausible, and empirically viable" theory as the basis for semantic annotation (Langacker, 2008), this project resulted in outcomes ranging from the mathematical formulae of the HCOMET metrics to the understanding required to improve HCOMET in future iterations. The process of training annotators, processing the corpus, and analyzing the data, allowed for both the conceptual and technical aspects of HCOMET to be scrutinized.

The specific outcomes of this project are listed below as well as the conceptual improvements envisioned for future iterations.

6.1 Outcomes

6.1.1 HCOMET, LEAF, SCENE: Three New Metrics Defined

The primary outcome of this project was the development of HCOMET, a cognitively and linguistically-informed human metric of semantic preservation in machine translation. In the process, two other scores, LEAF and SCENE, which determine more specific aspects of translation quality, were also constructed. The combination of these three scores may provide a better view into the semantic repercussions of machine translation systems.

6.1.2 Reliability Analysis

While the scoring functions conceptually measured semantic adequacy more accurately than previous attempts, the overall IAA of HCOMET did not surpass that of HMEANT. However, it seems that IAA can be increased through further training and individual experience. These results indicate that future studies would benefit from increasing the tutorial time and focusing on the aspects of UCCA whose IAA were lowest such as Grounds (G) and States (S).

6.1.3 Annotation Tool

The online and language-independent HCOMET tool for the alignment of parallel text was created. With this tool, future projects and iterations of HCOMET may continue to collect corpora. The UCCA annotation application was also updated by Omri Abend to include the last version of the UCCA annotation guidelines.

6.1.4 Parallel Corpora

The UCCA-annotated parallel corpus and its HCOMET equivalent corpus were collected and saved for possible future linguistic and statistical analyses. The UCCA corpus is available as XMLs while the HCOMET equivalent corpus is available as plain text in JSON format.

6.1.5 Better Understanding

Better understanding of the implications, linguistic as well as technical, of employing HCOMET as a metric of semantic preservation in machine translation is one of the most important outcomes of this project, given its exploratory nature. Thus, by creating the alignment application and analyzing the corpus gathered, ideas for future improvements were compiled and listed below:

1. The two-hour training session seemed to insufficiently explain UCCA and HCOMET according to several annotators.

2. The average time taken to annotate a single reference and machine translation pair ranged from 7 minutes to 49 minutes depending on the annotator. Perhaps more training and experience with UCCA and HCOMET could decrease the time taken by the slower annotators.

3. It seems that the current UCCA application and its annotation guidelines are not as easily applied to German as to English.

4. Annotators were not able to properly annotate some unintelligible machine translation output with the current UCCA application.

6.2 Future Research

From this point, HCOMET can be further developed in several directions. The new understanding of cognitive grammar as a plausible theory behind metrics of semantic preservation can be used to better HCOMET for future iterations. HCOMET may also be modified to account for bilingual semantic analyses. Furthermore, HCOMET may be reengineered as an automatic metric accounting for the cognitive and linguistic concepts behind annotations and alignments.

6.2.1 Improvements to HCOMET for Future Iterations

Incoherent Machine Translations

While the UCCA annotation tool was originally designed to process natural language, it was modified to also annotate machine translation output. Therefore, several of the checks and assertions required by the application were voided by the incoherent machine translated text encountered by annotators. Thus, future UCCA annotation tools should include specified methods to declare when translated text is incoherent to the point of being impossible to analyze through UCCA scenes.

Improved Training

It was shown that annotators understood some aspects of UCCA and HCOMET better than others. Given the low IAA of Grounds (G), States (S), and Adverbials (D), future training session should provide more examples and better definitions of these types of nodes.

Combining UCCA types

A possible method to increase IAA in HCOMET is to reclassify UCCA nodes into fewer categories. Through the process of annotation, it became clear that two separate classes of nodes were actually equivalent to one another. Scene level categories and sub-scene level categories were actually conceptually equivalent. Furthermore, the linguistic difference between certain semantically void prepositions marked as Functions (F) and others marked as Relators (R) is unclear at best. Perhaps, these two categories should be combined in order to avoid additional confusion. Future testing should determine if a re-categorization as proposed by the equivalences shown in Table 6.1 would increase HCOMET's

IAA.

Level Dependent Equivalences		
Scene Level		Sub-Scene Level
Adverbial (D)	$\approx$	Elaborator (E)
Linker (L)	$\approx$	Connector (N)
Other Equivalences		
Function (F)	$\approx$	Relator (R)

Table 6.1: Equivalence of UCCA Types

Introducing Semantic Roles

Although future HCOMET and UCCA iterations should still attempt to avoid using a complete set of semantic role labels, it would perhaps be useful to determine at least the agent and direct patient of each scene. Combining all participants under one label might not penalize systems enough, since sentences such as *He told the press...* and *The press said...* contain the same UCCA Participant (A) *the press*. However, the Participant fulfills a very different role in each of these sentences and should perhaps be penalized more than a simple partial alignment at the Scene level.

Introducing Types of Alignment

While the current definition of alignment in HCOMET is relatively vague, future iterations of the experiment may specify three different types of alignment to allow for more distinctive alignment analyses.

Unbound Alignment would determine if a semantic unit was translated intelligibly.

Scene Alignment would determine if a semantic unit was translated within the appropriate scene.

Role Alignment would determine if a semantic unit performs the appropriate function within its appropriate scene.

Accounting for Function Redunducy

As mentioned in Section 5.1.2, the conversion from UCCA to HCOMET may introduce redunduncies, which may affect the appropriate weighing of each HCOMET node. These discrepancies must be accounted for in future iterations.

Many-to-Many Alignment

Given that translation may not always retain the same number of scenes, future versions of HCOMET may allow many-to-many alignments so that partially-aligned nodes are not given a score of zero if they share the same alignment as described in Section 5.1.2.

6.2.2 Bilingual HCOMET

Since UCCA demonstrates "portability across domains and languages" (Abend and Rappoport, 2013), it is also possible to employ HCOMET in comparing a translation directly to its source in the original language. Such a method of semantic analysis would bypass the need for reference translations, perhaps reducing the cost of annotation depending on the availability of bilingual annotators. Reference translations may also not adequately convey cognitive metaphors or other aspects the original document, allowing bilingual HCOMET to also bypass any such translation issues.

6.2.3 From UCCA towards an Automatic Metric

While further research must be done before HCOMET can be fully automized into an automatic metric of semantic evaluation, simple methods may be used to automize the HCOMET alignment process to decrease annotation time. The alignment of leaf-nodes may be done automatically with any alignment toolkit such as Giza++ (Och and Ney, 2003). Once aligned, applications could also automatically determine the scenes in which each leaf node belongs, based on the UCCA tree annotated. Humans annotators would only need to determine whether the automatic alignments provided were accurate enough to warrant a full or partial alignment or whether they fulfilled the same semantic role within the scene. Such improvements may drastically decrease alignment time allowing for the creation of larger corpora for further statistical analyses.

Bibliography

Abend, O. and Rappoport, A. (2013). Universal conceptual cognitive annotation (ucca). In *Proceedings of ACL.* (on pp. i, 5, 6, 7, 8, 12, 13, 14, 15, 23, and 45)

ALPAC (1966). *Language and Machines: Computers in Translation and Linguistics*. The National Academies Press. (on p. 1)

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. (on p. 22)

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. (on pp. 1, 3, 11, 24, 25, 26, 27, and 28)

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44. (on pp. 1, 2, and 18)

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106. Association for Computational Linguistics. (on p. 2)

Cresswell, M. J. (1982). The autonomy of semantics. In *Processes, Beliefs, and Questions*, pages 69–86. Springer. (on p. 19)

Dixon, R. (2010a). *Basic Linguistic Theory Volume 1: Methodology*. Basic Linguistic Theory. OUP Oxford. (on p. 5)

Dixon, R. (2010b). *Basic Linguistic Theory Volume 2: Grammatical Topics*. Basic Linguistic Theory. OUP Oxford. (on p. 5)

Dixon, R. (2012). *Basic Linguistic Theory Volume 3: Further Grammatical Topics*. Basic Linguistic Theory. OUP Oxford. (on p. 5)

Dorr, B., Olive, J., McCary, J., and Christianson, C. (2011). Machine translation evaluation and optimization. In Olive, J., Christianson, C., and McCary, J., editors, *Handbook of Natural Language Processing and Machine Translation*, pages 745–843. Springer New York. (on p. 2)

Ge, R. and Mooney, R. J. (2009). Learning a compositional semantic parser using an existing syntactic parser. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 611–619. Association for Computational Linguistics. (on p. 20)

Koehn, P. (2009). A web-based interactive computer aided translation tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pages 17–20. Association for Computational Linguistics. (on p. 1)

Lakoff, G. (1972). *Linguistics and natural logic*. Springer. (on p. 19)

Langacker, R. W. (2008). *Cognitive Grammar : An Introduction: An Introduction*. Oxford University Press, USA. (on pp. 5, 25, and 41)

Lo, C.-k. and Wu, D. (2011). Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 220–229, Stroudsburg, PA, USA. Association for Computational Linguistics. (on pp. i, 3, and 24)

Mitchell, J. and Lapata, M. (2009). Language models based on semantic composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 430–439, Stroudsburg, PA, USA. Association for Computational Linguistics. (on p. 20)

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51. (on p. 45)

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics. (on pp. 1, 2, and 18)

Partee, B. (1995). Lexical semantics and compositionality. *An invitation to cognitive science: Language*, 1:311–360. (on p. 20)

Pinker, S. (1994). *The language instinct: The new science of language and mind*, volume 7529. Penguin UK. (on p. 20)

Przybocki, M., Sanders, G., and Le, A. (2006). Edit distance: a metric for machine translation evaluation. In *Proceedings of LREC-2006: fifth international conference on language resources and evaluation, Genoa, Italy*, pages 2038–2043. (on p. 2)

Sennrich, R. (2014). Bleualign/score.py. `https://github.com/rsennrich/Bleualign/blob/master/score.py`. (on p. 20)

Turian, J., Shen, L., and Melamed, I. D. (2003). Evaluation of machine translation and its evaluation. In *In Proceedings of MT Summit IX*, pages 386–393. (on p. 2)

Appendix A

HCOMET Instructions

A.1 Introduction

Goal:

To calculate how much meaning is preserved through translation.

Starting Points:

- Some translations are better than others.
- Machine translations can be difficult to understand.
- It is difficult to define exactly how good a translation is.

HCOMET

- HCOMET uses UCCA to divide the sentences into smaller units called nodes.
- It is easier to tell if these smaller units have been translated well.

A.2 What to Align

After annotating the reference and translation in UCCA, the software will automatically convert the annotations into two visual trees of HCOMET nodes. Some UCCA nodes will be automatically removed because they lack semantic content (e.g. function nodes). Your next step will be to align these HCOMET nodes.

A.2.1 Simple alignment steps

1. Begin with leaf (smallest) nodes.
    (a) It might be easier to start from the reference side.
    (b) For each leaf node, find its respective translation (if possible) and align.
    (c) If aligned, define whether they are partially the same or completely the same.
2. Begin to climb up the tree and align larger nodes.
    (a) Again, it might be easier to start from the reference side.
    (b) For each (larger) phrase find the corresponding translation node that contains the equivalent or corresponding information.
        - Prioritize nodes that share the same centers, processes, states, or participants, depending on which is the most important (or head) of the node.

A.2.2 Examples

Let’s compare the following four sentences:

1. John took a bath.
2. John had a shower.
3. John showered.
4. Took a bath John.

These sentences would be annotated in UCCA as follows:

1. John$_A$ [took$_F$ [a$_E$ bath$_C$]$_C$]$_P$.
2. John$_A$ [had$_F$ [a$_E$ shower$_C$]$_C$]$_P$.
3. John$_A$ showered$_P$.
4. [Took$_F$ [a$_E$ bath$_C$]$_C$]$_P$ John$_A$.

These sentences, in HCOMET, would have the following nodes[1]:

1. [John$_A$ [took [a$_E$ bath$_C$]$_C$]$_P$.]$_{Full}$
2. [John$_A$ [had [a$_E$ shower$_C$]$_C$]$_P$.]$_{Full}$
3. [John$_A$ showered$_P$.]$_{Full}$
4. [[Took [a$_E$ bath$_C$]$_C$]$_P$ John$_A$.]$_{Full}$

Let’s assume that Sentence 1 is the reference and all others are machine translations. Beginning with the smaller nodes and moving up, the following would be the alignments:

Leaf Nodes:
(1$_A$ refers to the node of type A in sentence #1)

1$_A$: John$_A$
= 2$_A$: John$_A$
= 3$_A$: John$_A$
= 4$_A$: John$_A$

1$_E$: a$_E$
= 2$_E$: a$_E$
= 4$_E$: a$_E$

1$_C$: bath$_C$
$\approx$ 2$_C$: shower$_C$
= 4$_C$: bath$_C$

Compound Nodes:

1$_C$: [a bath]$_C$
$\approx$ 2$_C$: [a shower]$_A$
= 4$_C$: [a bath]$_A$

1$_P$: [took a bath]$_P$
$\approx$ 2$_P$: [had a shower]$_P$
$\approx$ 3$_P$: [showered]$_P$[2]
= 4$_P$: [took a bath]$_P$

[1] Each underscript is the category of the node

[2] “bathed” would be a complete alignment.

Full Node:

1_{Full} : [John took a bath.]$_{Full}$

$\approx 2_{Full}$: [John had a shower.]$_{Full}$

$\approx 3_{Full}$: [John showered.]$_{Full}$

$\approx$[3] 4_{Full} : [Took a bath John.]$_{Full}$

A.2.3 Things to keep in mind

1. A node may be completely aligned to another node even though it does not include the same internal structure (e.g. [took a shower] and [showered]).

2. A sentence may have all of its corresponding parts correctly translated, but in a strange or incoherent order resulting in the *Full* sentence node being only partially aligned or not aligned at all, if the order makes the sentence not understandable.

3. If the translation in the example above were "John took a book," its smaller nodes of [book], [a book], and [took a book] would not be aligned. This is because the heads of those nodes would not even be similar to their corresponding reference nodes (e.g. [bath], [a bath], and [took a bath]). See alignment step 2b.

4. UCCA does not currently take into account tenses, meaning that "John is showering" and "John showered" would have the same UCCA annotations[4]. However, in HCOMET we should consider mistranslated tenses as partial alignments (if everything else is correct).

[3] Even though both the *P* and the *A* were completely aligned, their ordering was wrong, resulting in only a partial alignment for the full sentence. If the sentence weren't understandable it would not be aligned at all.

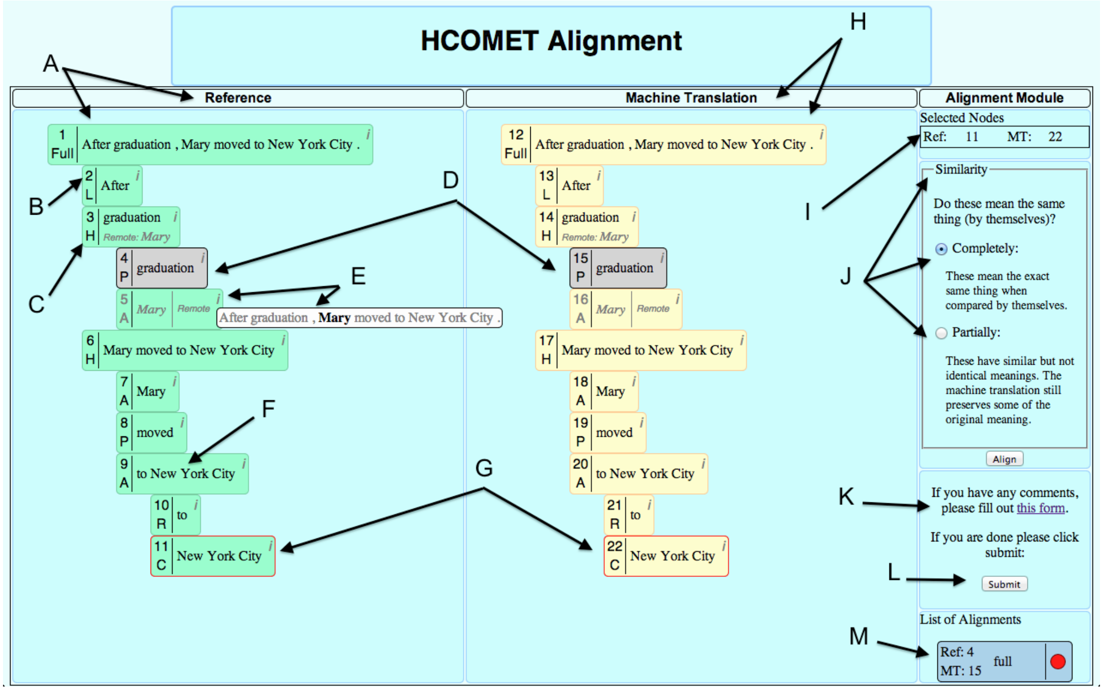
[4] [John]$_A$[is$_F$ showering$_C$]$_P$ has the same UCCA annotation as [John]$_A$[showered]$_P$

A.3 How to Align using the Web Application

List of Steps

1. Create the alignment by doing the following (in any order):
   - Click on the Reference Node you wish to align.
   - Click on the Translation Node you wish to align.
   - Click the appropriate button to define whether the alignment is Partial or Complete.
2. Click the “Align” button.
   - To delete an alignment, double click it on the list of alignments.
3. Repeat until done.
4. Click the “Submit” button when you are done. Click “OK” on the popup if you are sure.
5. After submitting, you will be redirected to the homepage.

Web Application Alignment Module Details:



Selected Nodes
Ref: 11 MT: 22

Similarity

Do these mean the same thing (by themselves)?

● Completely:

These mean the exact same thing when compared by themselves.

○ Partially:

These have similar but not identical meanings. The machine translation still preserves some of the original meaning.

If you have any comments, please fill out this form.

If you are done please click submit:

List of Alignments

Ref: 4 MT: 15 full ●

A Reference Translation

The reference translation will be displayed here.

B Node ID Number

This ID number is used to refer to this specific node (see F).

C UCCA Type

The UCCA type of this node in context can be found here.

D Aligned Nodes

Nodes that have already been aligned will be greyed-out and no longer selectable.

E Text in Context

To see the text of this node in the context of the whole sentence, hover your mouse over the greyed-out *i*.

F Text

This node contains the text found here.

G Selected Nodes

To select the nodes you want to align, simply click them and they will be outlines and appear in the Selected Nodes box (see I).

H Machine Translation

The machine translation will be displayed here.

I Selected Nodes Box

The selected nodes (see G) will be displayed here.

J Similarity

After choosing the nodes to be aligned, define whether they are completely or partially aligned (as instructed).

K Comments Form

If there is something particularly interesting about this sentence pair, please click on the link and fill out the google form to submit your comments.

L Submit

Once the alignment is complete, click submit and a popup should appear asking whether you are truly done. If you are, click OK and you will be redirected to the home screen.

M List of Alignments

After each you click align, a new alignment will be created and will be displayed. To delete an alignment, simply double click it.

A.4 More Complex Examples

Notation

In the following examples:

The p*A*rticipant of the reference is referred to as Reference$_A$.

The *P*rocess of the translation will be referred to as Translation$_P$.

Etc...

Remember that the software will automatically convert between the UCCA annotations you entered and the HCOMET nodes as the following examples. Certain UCCA nodes will be removed before the alignment step but must still be annotated.

A.4.1 Active & Passive

Reference

Text: Elizabeth seems to have kicked the soccer ball.

UCCA[5]: [Elizabeth]$_A$ [seems]$_D$ [to$_F$ have$_F$ kicked$_C$]$_P$ [the$_E$ soccer$_E$ ball$_C$]$_A$.

HCOMET: [[Elizabeth]$_A$ [seems]$_D$ [to have [kicked]$_C$]$_P$ [the$_E$ soccer$_E$ ball$_C$]$_A$.]$_{Full}$

Translation

Text:The soccer ball seems to have been kicked by Elizabeth.

[5] UCCA annotations will be shown as footnotes after this one.

HCOMET[6]:[[The$_E$ soccer$_E$ ball$_C$]$_A$ [seems]$_D$ [to have been [kicked]$_C$]$_P$ [by$_R$ Elizabeth$_C$]$_A$.]$_{Full}$

Even though these sentences are different, their main components are completely aligned:

Reference$_A$: [Elizabeth]$_A$
= Translation$_A$: [by Elizabeth]$_A$

Reference$_P$: [seems to have kicked]$_P$
= Translation$_P$: [seems to have been kicked]$_P$

Reference$_A$: [the soccer ball]$_A$
= Translation$_A$: [The soccer ball]$_A$

Notes:

The passive sentence includes the UCCA function node [been]$_F$, but function nodes are not shown in HCOMET, therefore they have no effect on whether the full meaning has been maintained.

[6] UCCA: [The$_E$ soccer$_E$ ball$_C$]$_A$ [seems]$_D$ [to$_F$ have$_F$ been$_F$ kicked$_C$]$_P$ [by$_R$ Elizabeth$_C$]$_A$.

A.4.2 Errors in the Leaf Nodes

Reference

Text: John studies media with a focus on advertising.

HCOMET[7]: [John$_A$ studies$_P$ [media$_C$ [[with a focus$_C$ on]$_R$ advertising$_C$]$_E$]$_A$.]$_{Full}$

Translation

Text: John is studying media with a focus on broadcasting.

HCOMET[8]: [John$_A$ [is studying$_C$]$_P$ [media$_C$ [[with a focus$_C$ on]$_R$ broadcasting$_C$]$_E$]$_A$.]$_{Full}$

As you may have noticed, the only translation error is the leaf node [broadcasting]$_C$. Because this node is within many larger nodes, this error will affect all of their translation similarities. The following list is not exhaustive, but contains all of the nodes affected.

Reference$_C$: [advertising]$_C$

$\neq$ Translation$_C$: [broadcasting]$_C$

Reference$_E$: [with a focus on [adversiting]]$_E$

$\approx$ Translation$_E$: [with a focus on [broadcasting]]$_E$

Reference$_A$: [media [with a focus on adversiting]]$_A$

$\approx$ Translation$_A$: [media [with a focus on broadcasting]]$_A$

Reference$_{Full}$: [John is studying [media with a focus on adversiting]]$_{Full}$

$\approx$ Translation$_{Full}$: [John is studying [media with a focus on broadcasting]]$_{Full}$

Notes:

The translation errors smaller nodes are carried through to the larger nodes.

[7] UCCA: John$_A$ studies$_P$ [media$_C$ [[with$_F$ a$_F$ focus$_C$ on$_F$]$_R$ advertising$_C$]$_E$]$_A$.

[8] UCCA: John$_A$ [is$_F$ studying$_C$]$_P$ [media$_C$ [[with$_F$ a$_F$ focus$_C$ on$_F$]$_R$ broadcasting$_C$]$_E$]$_A$.

A.4.3 Errors in Ordering

Reference

Text: John let Mary down.

HCOMET[9]: [John$_A$ [let...[10]]$_{P_{(start)}}$ Mary$_A$ [... down]$_{P_{(cont.)}}$.]$_{Full}$

Translation

Text: Mary disappointed John.

HCOMET[11]: [Mary$_A$ disappointed$_P$ John$_A$.]$_{Full}$

The error in this sentence pair is in the order of the participants. Each of the leaf nodes is correctly translated and each of the compound nodes is also correctly translated.The only node that is affected is the $Full$ node.

These are some of the nodes that are completely aligned:

Reference$_A$: [Mary]$_A$

= Translation$_A$: [Mary]$_A$

Reference$_P$: [let ... down]$_P$

= Translation$_P$: [disappointed]$_P$

All other leaf and compound nodes...

This is the $Full$ Node that is only partially aligned:

Reference$_{Full}$: [John let Mary down.]$_{Full}$

$\approx$ Translation$_{Full}$: [Mary disappointed John.]$_{Full}$

Notes:

It is important to make sure all of the leaf nodeas are correctly aligned as well as the $Full$ Node. Even if all of the smaller nodes are correctly aligned, the larger node may have ordering errors.

[9] UCCA: John$_A$ [let...]$_{P_{(start)}}$ Mary$_A$ [... down]$_{P_{(cont.)}}$.

[10] The node [forced ... let down]$_P$ is a single node but is discontinuous.

[11] UCCA: Mary$_A$ disappointed$_P$ John$_A$.

Appendix B

HCOMET Quick Guide

After annotating the reference and translation in UCCA, the software will automatically convert the annotations into two visual trees of HCOMET nodes. Some UCCA nodes will be automatically removed because they lack semantic content (e.g. function nodes). Your next step will be to align these HCOMET nodes.

Aligning Nodes

1. Begin with smallest nodes.
   (a) It might be easier to start from the reference side.
   (b) For each node, find its respective translation (if possible) and align.
   (c) If aligned, define whether they are completely equivalent or just partially equivalent.
2. Begin to climb up the tree and align larger nodes.
   (a) Again, it might be easier to start from the reference side.
   (b) For larger phrases find the corresponding translation node that contains the corresponding information.
      - Prioritize nodes that share the same centers, processes, states, or participants, depending on which is the most important (i.e. the head of the node).

Things to keep in mind

1. A node may be completely aligned to another node even though it does not include the same internal structure (e.g. [took a shower] and [showered]).

2. A sentence may have all of its corresponding parts correctly translated, but in a strange or incoherent order resulting in the $Full$ sentence node being only partially aligned or not aligned at all, if the order makes the sentence not understandable.

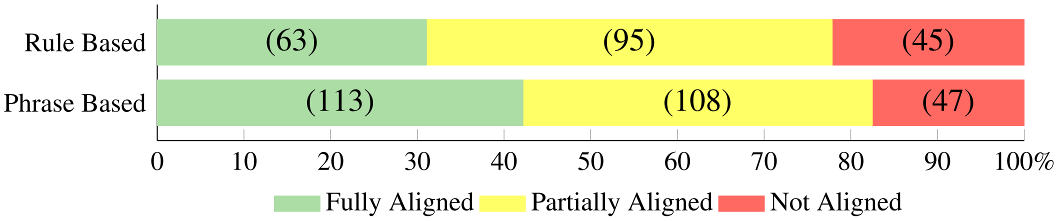
3. If the translation in the example above were “John took a book,” its smaller nodes of [book], [a book], and [took a book] would not be aligned. This is because the heads of those nodes would not even be similar to their corresponding reference nodes (e.g. [bath], [a bath], and [took a bath]). See alignment step 2b.

4. UCCA does not currently take into account tenses, meaning that “John is showering” and “John showered” would have the same UCCA annotations[1]. However, in HCOMET we should consider mistranslated tenses as partial alignments (if everything else is correct).

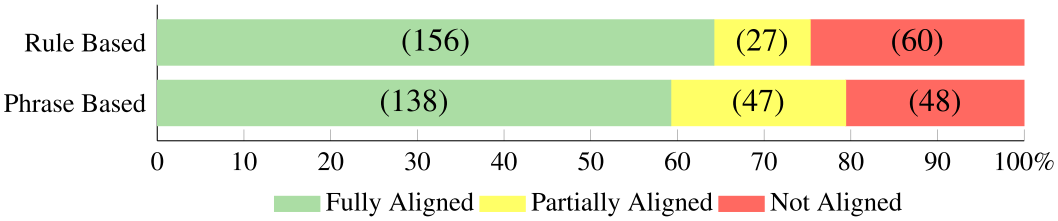
[1] [John]$_A$[is$_F$ showering$_C$]$_P$ has the same UCCA annotation as [John]$_A$[showered]$_P$

Appendix C

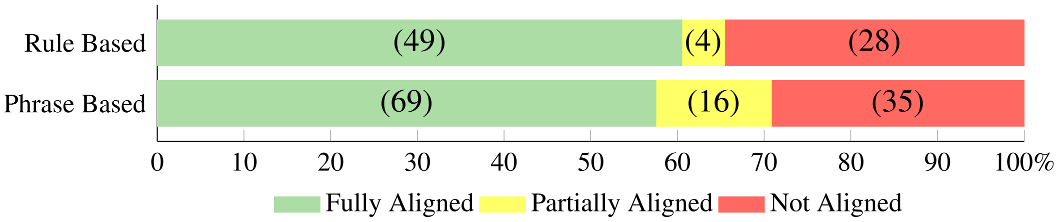
Alignments of German to English Machine Translations



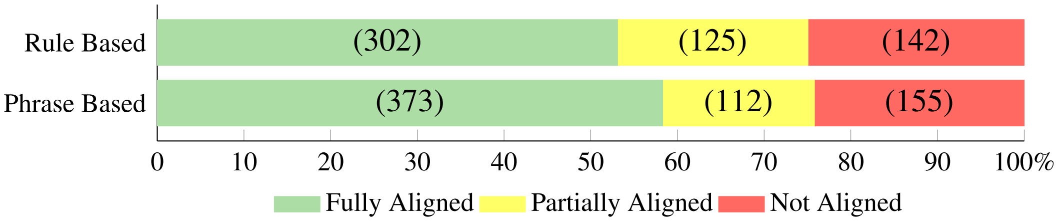
Alignment of Parallel Scenes (H): Proportion and (Count)



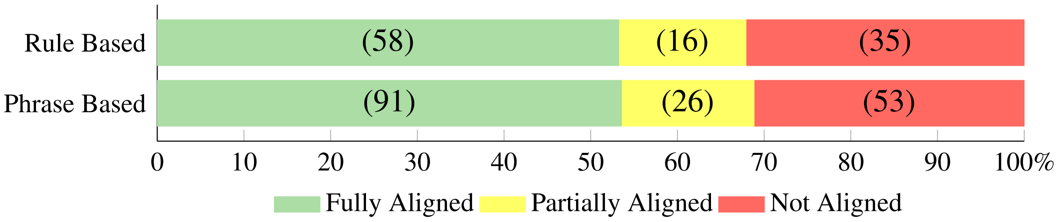
Alignment of Processes (P): Proportion and (Count)



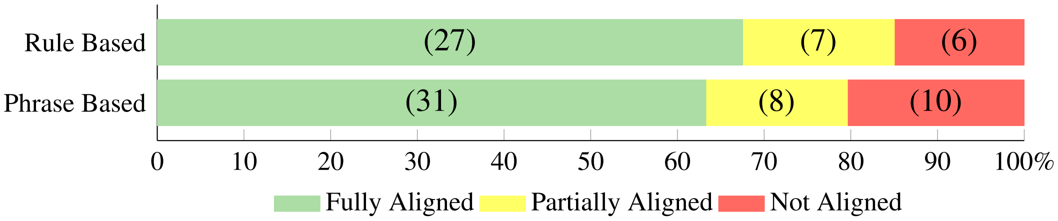
Alignment of States (S): Proportion and (Count)



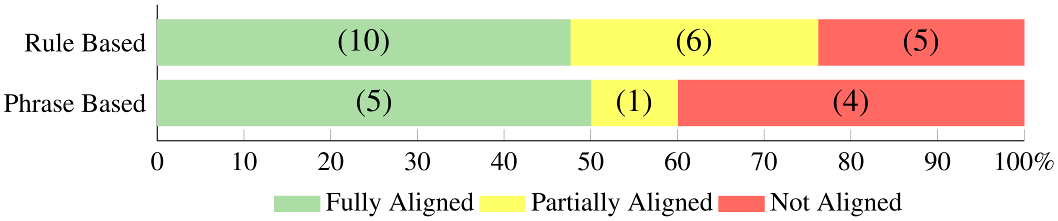
Alignment of Participants (A): Proportion and (Count)



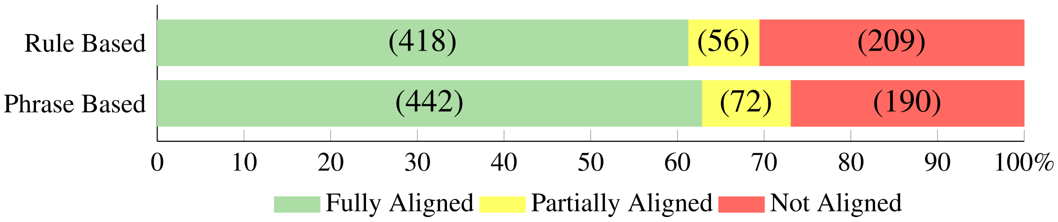
Alignment of Adverbials (D): Proportion and (Count)



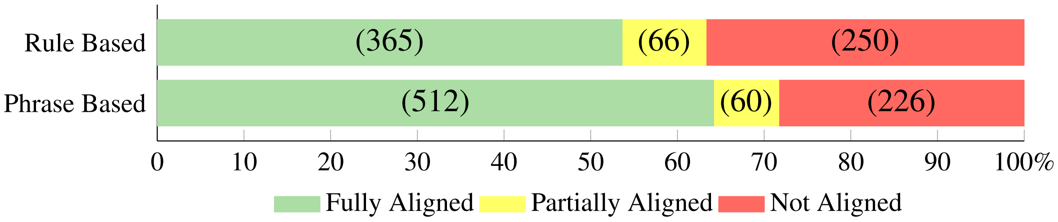
Alignment of Time Words (M): Proportion and (Count)



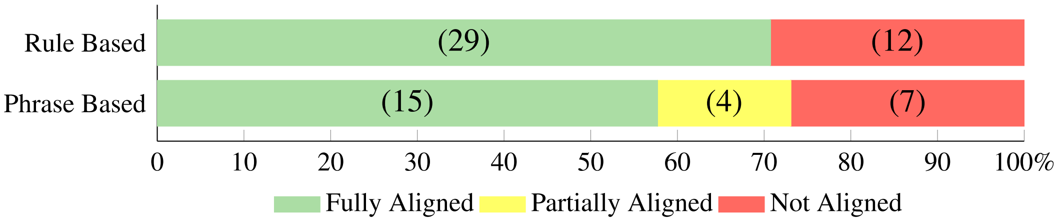
Alignment of Grounds (G): Proportion and (Count)



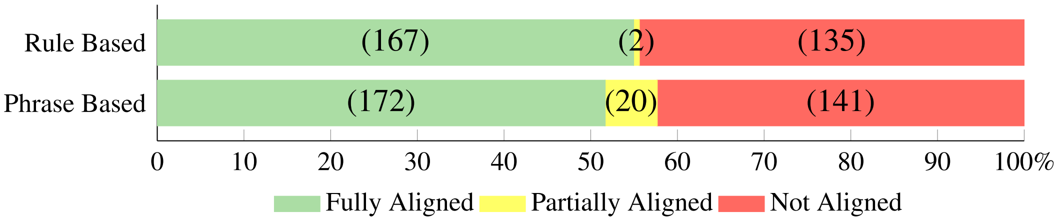
Alignment of Centers (C): Proportion and (Count)



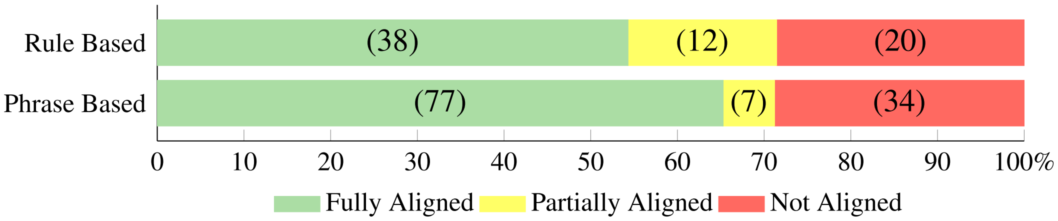
Alignment of Elaborators (E): Proportion and (Count)



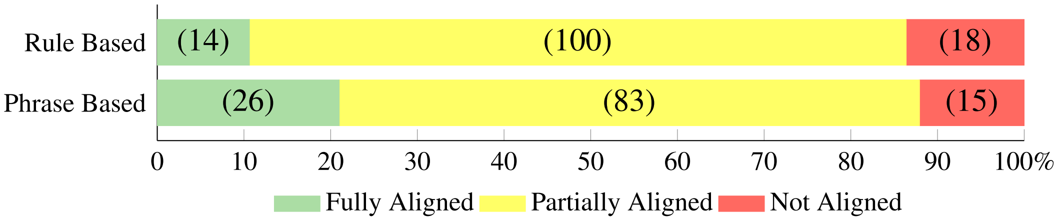
Alignment of Connectors (N): Proportion and (Count)



Alignment of Relators (R): Proportion and (Count)



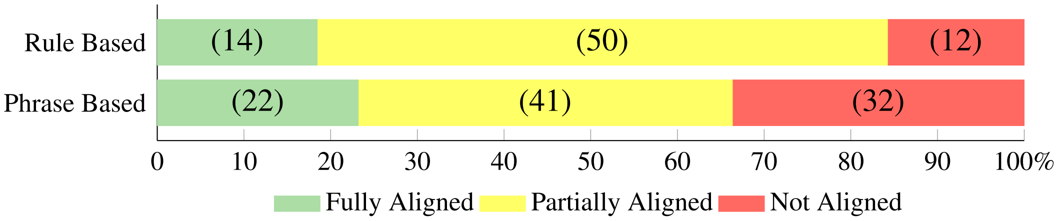
Alignment of Linkers (L): Proportion and (Count)



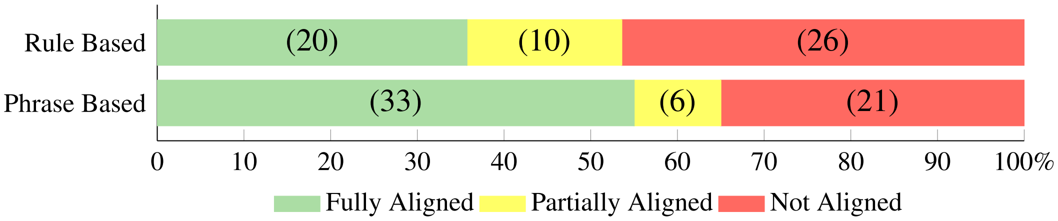
Alignment of Full Sentences (Full): Proportion and (Count)

Appendix D

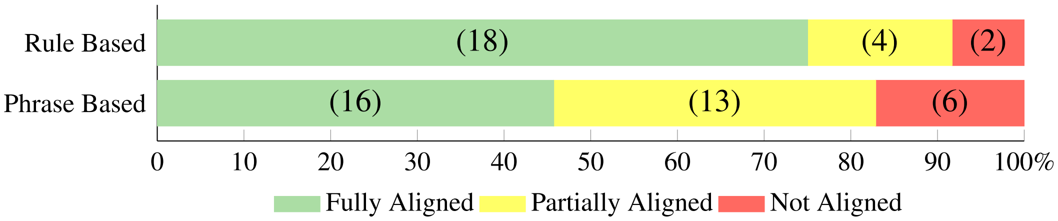
Alignments of English to German Machine Translations



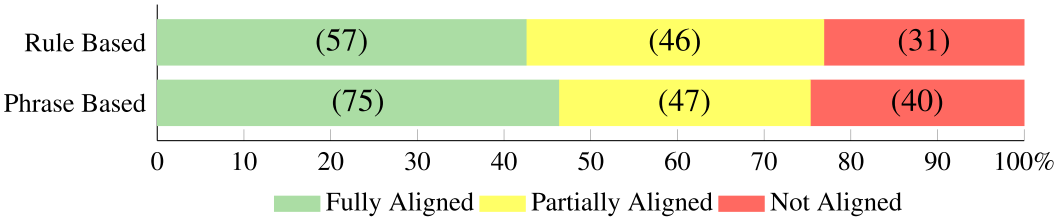
Alignment of Parallel Scenes (H): Proportion and (Count)



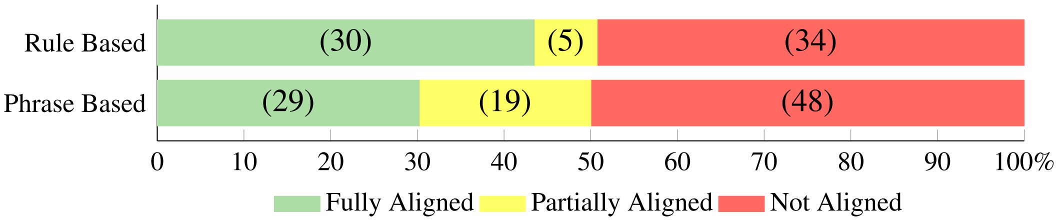
Alignment of Processes (P): Proportion and (Count)



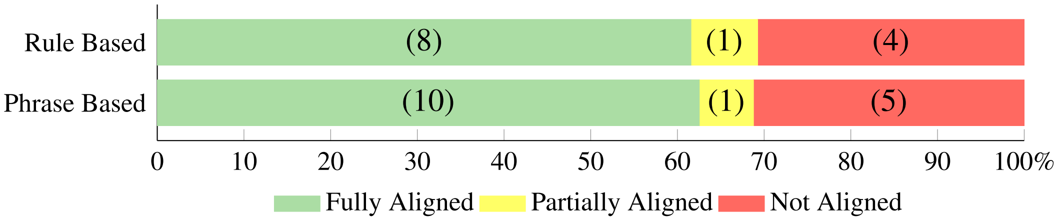
Alignment of States (S): Proportion and (Count)



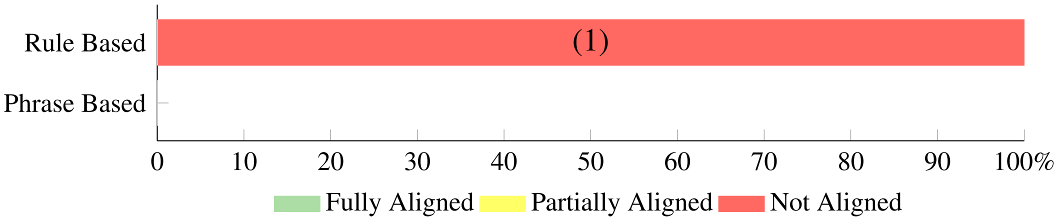
Alignment of Participants (A): Proportion and (Count)



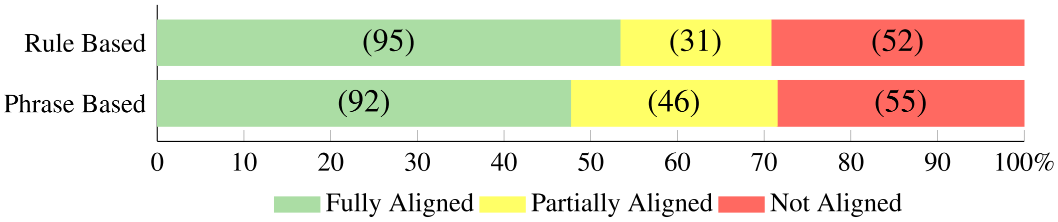
Alignment of Adverbials (D): Proportion and (Count)



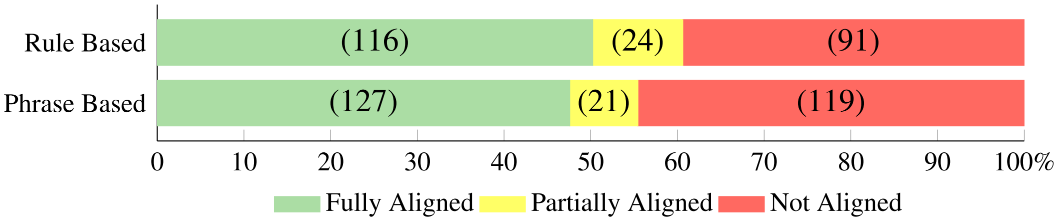
Alignment of Time Words (M): Proportion and (Count)



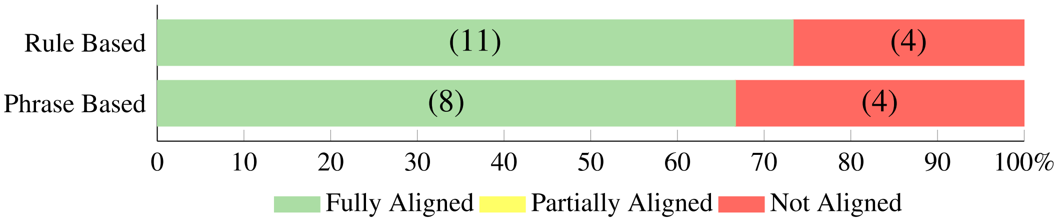
Alignment of Grounds (G): Proportion and (Count)



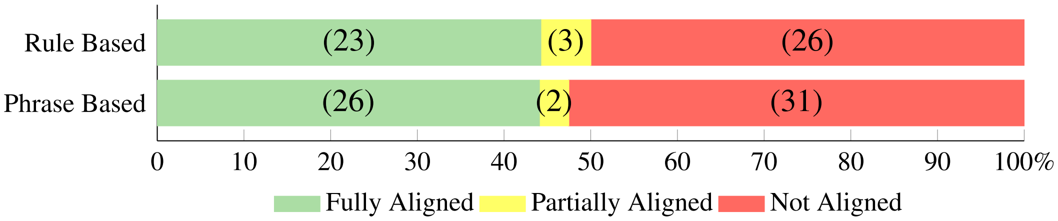
Alignment of Centers (C): Proportion and (Count)



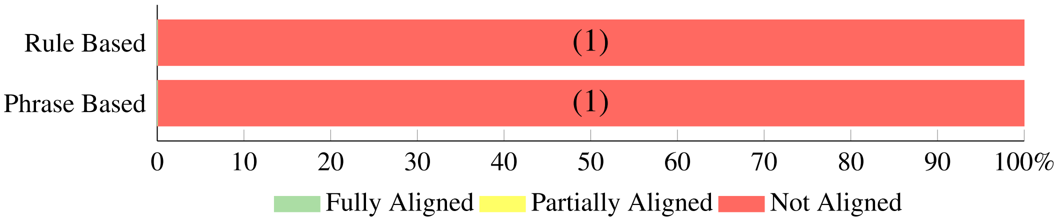
Alignment of Elaborators (E): Proportion and (Count)



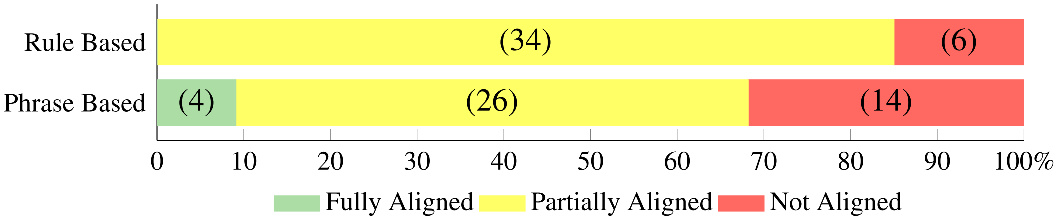
Alignment of Connectors (N): Proportion and (Count)



Alignment of Relators (R): Proportion and (Count)



Alignment of Linkers (L): Proportion and (Count)



Alignment of Full Sentences (Full): Proportion and (Count)