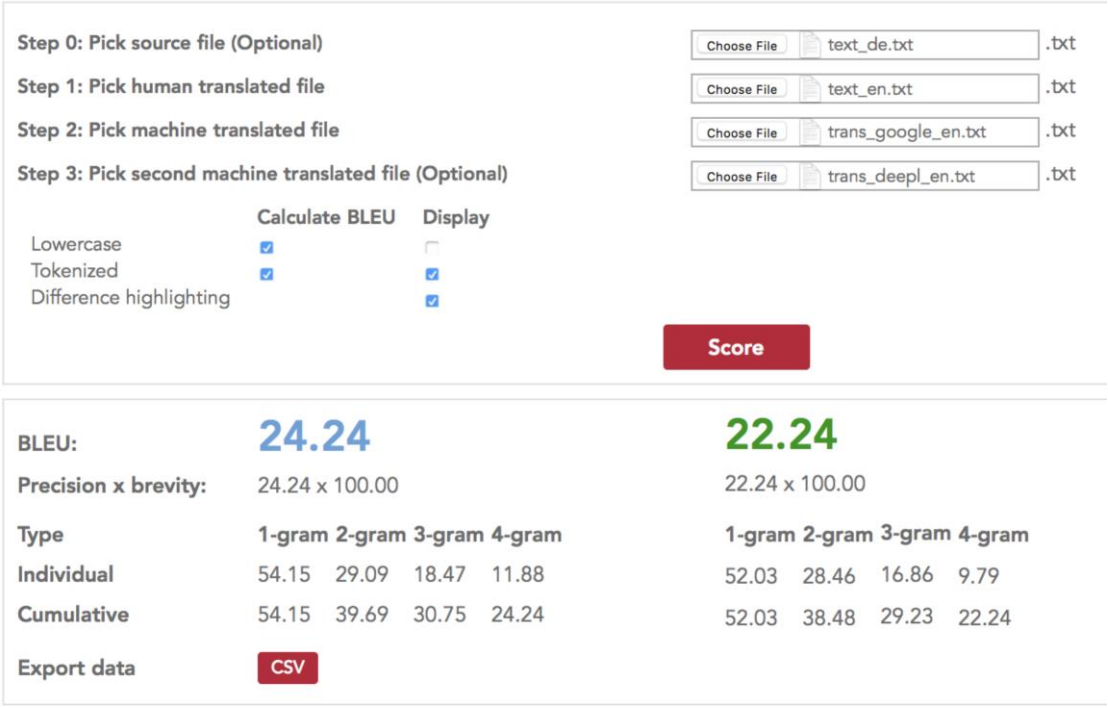
Assignment 1

Aufgabe 2.1

	DE → EN	EN → DE
Google Translate	0.271	0.214
DeepL	0.273	0.215
Bing	0.254	0.168

Aufgabe 2.2

Google Translate (blau) vs DeepL (grün) DE → EN



Die BLEU - Werte von Google Translate sind höher als von DeepL.

Aufgabe 2.3

Der BLEU-Wert von einem System ist höher, aber der Satz ist nach eurer Meinung schlechter übersetzt als beim anderen.

Sentence 4	BLEU	Length ratio	Text
Source	-	-	Demonstrieren sende womöglich das falsche Signal, sagt Student Johannes Müller.
Human	100.00	1.00	That might send the wrong signal , says student Johannes Müller .
Machine	81.55	0.92	Demonstrate send the wrong signal , says student Johannes Müller .
Machine	80.71	1.00	Demonstrating may send the wrong signal , says student Johannes Müller .

Wir glauben, dass DeepL (grün) den Satz trotz der niedrigen BLEU – Werte besser als GT (blau) übersetzt hat

Die Referenz Übersetzung ist nur eine sinngemässe Übersetzung, weswegen die BLEU- Werte der Systeme schlecht sind, obwohl die Übersetzungen gut sind.

Sentence 3	BLEU	Length ratio	Text
Source	-	-	"Wir müssen aktiv werden," sagt Dozentin Nina Hall.
Human	100.00	1.00	" We need to take action , " says professor Nina Hall .
Machine	22.63	1.00	" We have to become active , " says lecturer Nina Hall .
Machine	35.74	1.00	" We need to get active , " says lecturer Nina Hall .

Sentence 9	BLEU	Length ratio	Text
Source	-	-	Moment, sagen die anderen:
Human	100.00	1.00	Hold on , say others :
Machine	22.96	1.00	Wait , say the others :
Machine	22.96	1.00	Wait , say the others :

Eine Übersetzung, die euch amüsiert.

Sentence 7	BLEU	Length ratio	Text
Source	-	-	Seit Monaten bringen einige Daten in Sicherheit.
Human	100.00	1.00	For months , some of them have been saving data .
Machine	26.30	0.73	For months , some data to safety .
Machine	35.66	1.09	For months , some of the data has been kept safe .

Google Translate (blau) amüsiert uns.

Task 2.4

• Ihr wollt überprüfen, ob euer neues MU-System ein spezielles linguistisches Phänomen besser übersetzt als ein Standard-MU-System.

> Ein spezielles linguistisches Phänomen hat wahrscheinlich nicht besonders viele Testsätze, die man untersuchen kann. Folglich ist es einfacher, die Evaluation manuell vorzunehmen. Auch weil ein so spezifisches Phänomen ein wenig mehr Feinfühligkeit verlangt, als eine Maschine zu bieten hat (Kontrolle ob es richtig ist ist schwieriger).

• Ihr habt ein MU-System für eine Firma gebaut und müsst die Projektleiter über die Qualität des resultierenden Systems informieren.

> Hier würden wir eine automatische Evaluation verwenden. Dies, da für eine aussagekräftige Präsentation genug Testsätze ausgewertet werden müssen. Und hierfür einen Menschen anzustellen, kostet sehr viel Geld. Man könnte hier aber auch eine halbautomatische Evaluation verwenden, wenn die Firma genug Geld hat.

• Ihr organisiert einen Wettbewerb für Maschinelle Übersetzung und müsst viele Systeme vergleichen, um den Gewinner festzulegen.

> Auch hier wieder die automatische Evaluation. Es lohnt sich einfach nicht so viel Zeit und Geld in diese Auswertung zu stecken. Auch die Menge spielt hier eine Rolle. Wären es nur drei Teilnehmer mit je 2 Testsätzen, könnte man es auch gut von Hand machen.

• Ihr startet ein neues Projekt und möchtet zuerst einen Überblick über die Qualität verschiedener Systeme erhalten, um zu entscheiden, welches davon ihr weiter entwickeln wollt.

> Hier die manuelle Evaluation. Denn je nach Thematik des Projekts hat mein eine spezifische Domäne der Testsätze und es ist halt immer noch am genausten, dies von Hand zu bewerten. Ausserdem werden auch nicht tausende Sätze getestet.