Annotating translation errors in a machine translated novel

To be able to assess the quality of Google's Neural Machine Translation system on the translation of Agatha Christie's novel `The Mysterious Affair at Styles' into Dutch all errors in the MT output will be indicated and labelled.

To be able to judge the quality of the translations, they will be marked for two important aspects: accuracy and fluency. Accuracy is concerned with the relationship between source text and target text, whereas fluency is concerned with the target text and language.

You will use a two-step approach. In a first step only the translation (target sentence) will be visible and all fluency errors will be annotated; in a second step, the source sentence will be displayed as well and you have to compare the source and target sentence and indicate all accuracy errors. To label the errors, you will make use of a predefined error taxonomy.

Step 1: Fluency errors

The goal of the first step is to annotate the machine translated text for fluency. Fluency can be described as respecting the norms of the target language and culture. A good translation should read as a native Dutch text. This includes respecting the conventions of the language (grammar, lexicon, spelling) as well as respecting the conventions of the text structure (paragraph content and coherence) and the text type, in this case literature.

As an annotator, it is your task to mark anything that does not follow the conventions of the Dutch language or that does not respect the demands of the text structure or the text type. To facilitate the task, different categories (Coherence, Lexicon, Style & register, Grammar & Syntax and Spelling) are predefined in WebAnno, along with their most important subcategories.

Note: Annotators can neglect punctuation marks.

<u>Coherence</u>

There is something wrong with the coherence of the text: lack of logical structure, confusing relationships, lack of logical structure, …

> <u>Logical problem</u>: the information as such is illogical/confusing/makes no sense when looking at the rest of the text

We hadden **Logical problem I Coherence** een goed garen over oude tijden en het eindigde met zijn uitnodiging om naar Styles te gaan om daar mijn verlof door te brengen.

"De **Logical problem I Coherence** mater zal blij zijn om je weer te zien ... na al die jaren," voegde hij eraan toe.

"Ik heb **Logical problem I Coherence** een neef die borstvoeding geeft," merkte ik op.

Non-existing word: the word does not exist in Dutch and makes no sense

Mevr. Cavendish was echter een dame die graag haar eigen plannen maakte, en verwachtte

dat andere mensen bij **Logical problem I Coherence** hen zouden vallen, en in dit geval had ze zeker de

Non-existing word I Coherence zweephand , namelijk: de **Non-existing word I Coherence** buidelrnen .

Cultural reference: the word/phrase refers to something specific of the culture of the source, but makes no sense in the target language.

Een jong meisje in **Cultural reference I Coherence** V.A. D. uniform liep licht over het gazon.

Discourse marker: the conjunction or linking word expresses a strange relationship

Discourse marker I Coherence Inderdaad , ik had hem **Discourse marker I Coherence** nog nooit zo goed gekend.

Hij was ruim vijftien jaar ouder dan ik, in de eerste plaats **Discourse marker I Coherence** omdat hij er amper vijfenveertig jaar uitzag.

Co-reference: co-reference problem, mismatch between entities, e.g. feminine pronoun to refer to a male person

Ik was **Logical problem I Coherence** buiten het front geïnvalideerd; en na een aantal maanden in een nogal deprimerend **Logical problem I Coherence** herstellend huis te hebben doorgebracht, kreeg **Co-reference I Coherence** hij een maand ziekteverlof.

Maar de **Logical problem I Coherence** materie was hem meteen **Logical problem I Coherence** aangegrepen , nam hem aan als **Co-reference I Coherence** secretaresse ... je weet hoe ze altijd honderd **Logical problem I Coherence** samenlevingen bestuurt? '

Inconsistency: a term or a notation is used inconsistently throughout the text

Het dorp Styles St. Mary lag ongeveer drie kilometer van het kleine station

Ze werkt in het Rode Kruis-ziekenhuis in Tadminster, zeven [Inconsistency I Coherence] mijl verderop.

[Style & Register I Disfluent sentence/construction]
" Waar is thee vandaag ... binnen of buiten?"

[Inconsistency I Coherence]
" Uit .

Verb tense: the tense of the verb is wrong/illogical in the context of the rest of the sentence/text

wilde, ongetemde geest in een prachtig geciviliseerd lichaam ... al deze dingen

[Verb tense I Coherence]
worden in mijn geheugen gebrand.

Lexicon

One lexical element does not entirely fit in the Dutch sentence.

Lexical choice: the meaning of the word is clear in the context, but it is not idiomatic in Dutch

Lawrence, de [Lexicon I Lexical choice] jongere , [Lexicon I Lexical choice] was (flu) een delicate jeugd [Lexicon I Lexical choice] geweest .

[Lexicon I Lexical choice]
Ze wierp zich naast John op de grond en terwijl ik haar een bord boterhammen overhandigde, glimlachte ze naar me.

[Lexicon I Lexical choice]
een grondig fascinerende vrouw.

Wrong preposition: this expression requires a different preposition

[Lexicon I Wrong preposition]
Ik zag meteen door zijn gezicht dat er iets verontrustends had plaatsgevonden

[Lexicon I Wrong preposition]
De intense interesse die in het publiek werd gewekt door wat destijds bekend stond als 'The Styles Case' is nu enigszins verdwenen.

Grammar & syntax

Anything that does not follow the grammatical or structural rules of the Dutch language

Agreement: The sentence is grammatically incorrect because there is a mismatch between and article and noun, subject-verb, etc.
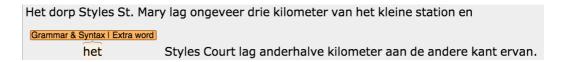
Grammar & Syntax | Agreement (flu)
Logical problem | Coherence   Grammar & Syntax | Agreement
Dit vertrouwen   zullen   de sensationele geruchten stoppen die nog steeds bestaan.

Verb form: a grammatically incorrect verb form

Word order: a grammatically incorrect word order

Grammar & Syntax | Word order (flu) Grammar & Syntax | Word order
"Geen geheime hobby?"   zij   vroeg   .

Extra words: The sentence is grammatically incorrect because there is a superfluous constituent or word. This can be a constituent that already appeared earlier in the sentence or a word that has been written twice in a row

Het dorp Styles St. Mary lag ongeveer drie kilometer van het kleine station en
Grammar & Syntax | Extra word
het   Styles Court lag anderhalve kilometer aan de andere kant ervan.

Missing words: The sentence is grammatically incorrect because a necessary part of the structure is missing, this can be a preposition or an entire constituent such as an obligatory direct object.

Grammar & Syntax | Missing word
Daarna   wendde zich   tot zijn vrouw:

Grammar & Syntax | Missing word
Miss Howard   schudde de   hand met een stevige, bijna pijnlijke greep.

Grammar & Syntax | Missing word
Ik sprak mijn waardering uit, en John bracht me naar het huis   en de   brede trap op, die rechts en links voorwaarts slingerde naar verschillende vleugels van het gebouw.

Style & register

Disfluent sentence/construction: The sentence / constituent is not grammatically incorrect, but it is nonetheless very difficult to read, it could be translated in a much more idiomatic way.

Ik had een paar jaar   heel weinig van hem gezien   .

Mevr. Cavendish, die met zijn vader was getrouwd toen hij weduwnaar met twee zonen was,

was een knappe vrouw van middelbare leeftijd geweest,   zoals ik haar herinnerde   .

Repetition: The same or a very similar word/expression is used more than once

en een knappe oude   oude   dame

Cynthia Murdoch was een jong uitziend   jong   wezen, vol leven en kracht.

Register: The words have the same meaning, but the chosen word/expression is too formal/informal/... for the text or belongs to a regional variety of the language that is not entirely suitable for the target audience.'

Untranslated: A word / fragment of which a Dutch translation exists is left untranslated.

"En ze is doodsbang voor '   Sisters   '."

Spelling

Spelling mistake: wrong spelling of a word

"Pesten ze je dan?" Vroeg ik glimlachend.

Een dame in een stevige tweedrok,

Other

Step 2: Accuracy errors

The goal of the second step is to annotate the machine translated text for accuracy. Accuracy can be described as making sure the target text contains the same information as the source text. This means that all misinterpretations, contradictions, meaning shifts, additions or deletions are potential errors.

1. Before you start annotating for accuracy, make sure the Accuracy layer is selected in Webanno.
2. Accuracy annotations require two sets of annotations (one in source sentence and one in target sentence) and a "link" from source to target annotation.
3. First annotate the source text error, then annotate the corresponding translation in target text. Link annotation in the source text to the annotation in the target text with an arrow.

Accuracy has 6 main categories to choose from: Mistranslation, Do Not Translate, Untranslated, Addition, Omission and Capitalization & Punctuation.
Only Mistranslation has subcategories, namely Multiword, Word sense, Semantically unrelated, Part of Speech, Partially translated.

Mistranslation

Source content has been translated (when it should be translated) but the translation I incorrect. Choose the suitable subcategory.

> Multiword: The translation is incorrect (and often too literal) because the English sentence contained a multi-word expression such as an idiom, a proverb, a collocation, a compound or a phrasal verb. Idiomatic expressions and proverbs are indicated with a paragraph sign (¶) in Van Dale.
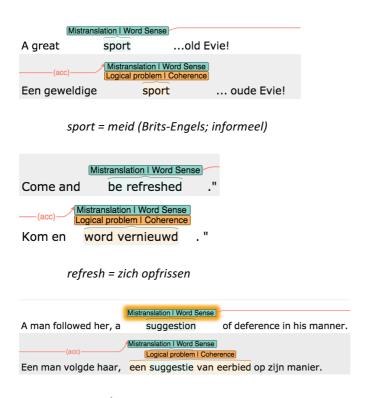

> to fall in with = zich aansluiten bij
> to have the whip hand = macht hebben over
> purse strings = touwtjes van de geldbeurs

Mistranslation | MWE (acc) **Mistranslation | MWE** (acc)

But you could have **knocked** us all **down with a feather** when, three months ago, she suddenly

(acc)

announced that she and Alfred were engaged!

Mistranslation | MWE
Logical problem | Coherence
(acc)

Maar je had ons allemaal **met een veer kunnen slaan** toen ze drie maanden geleden plotseling aankondigde dat zij en Alfred verloofd waren!

knock down with a feather = zeer verbaasd zijn

Word sense: The target content refers to a different (and a wrong) sense of the source content.

Mistranslation | Word Sense

A great **sport** ...old Evie!

Mistranslation | Word Sense
Logical problem | Coherence
(acc)

Een geweldige **sport** ... oude Evie!

sport = meid (Brits-Engels; informeel)

Mistranslation | Word Sense

Come and **be refreshed** ."

Mistranslation | Word Sense
Logical problem | Coherence
(acc)

Kom en **word vernieuwd** . "

refresh = zich opfrissen

Mistranslation | Word Sense

A man followed her, a **suggestion** of deference in his manner.

Mistranslation | Word Sense
Logical problem | Coherence
(acc)

Een man volgde haar, **een suggestie** van eerbied op zijn manier.

suggestion = spoor, zweem

Mistranslation | Word Sense (acc)

With the presence of Mr. Inglethorp, a sense of **constraint** and veiled hostility seemed to

(acc)
(acc)

Mistranslation | Word Sense

settle down upon the **company** .

(acc)
(acc)

Mistranslation | Word Sense
Logical problem | Coherence

Met de aanwezigheid van de heer Inglethorp leek een gevoel van **dwang** en verhulde

(acc)

Mistranslation | Word Sense
Logical problem | Coherence

vijandigheid zich te vestigen **op het bedrijf** .

constraint = verkrampt gedrag
company = gezelschap

Semantically unrelated: The meaning of one or more translated words is not related in any way to the meaning of the source word(s) and does not make any sense in the context.
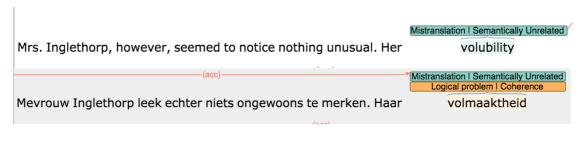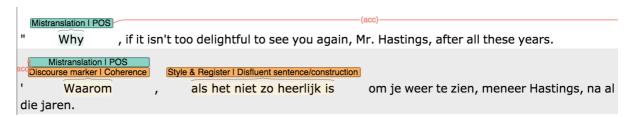
Mrs. Inglethorp, however, seemed to notice nothing unusual. Her

Mistranslation | Semantically Unrelated
volubility

(acc)

Mistranslation | Semantically Unrelated
Logical problem | Coherence

Mevrouw Inglethorp leek echter niets ongewoons te merken. Haar volmaaktheid

volubility = welbespraaktheid

Part-of-Speech: The translation represents an incorrect lexical category (Part-of-Speech) of the corresponding source text.

Mistranslation | POS   (acc)

" Why , if it isn't too delightful to see you again, Mr. Hastings, after all these years.

Mistranslation | POS
Discourse marker | Coherence   Style & Register | Disfluent sentence/construction

' Waarom , als het niet zo heerlijk is om je weer te zien, meneer Hastings, na al die jaren.

why = tussenwerpsel hier: wel

Partially translated: The translation is incorrect, e.g. due to the partial translation of a Dutch separable verb.

Mistranslation | Partial
John practiced for some time as a barrister,

Mistranslation | Other
congenial life of a country squire .

(acc)
Mistranslation | Partial
Logical problem | Coherence
John oefende enige tijd als advocaat,

practice = een beroep uitoefenen

> Mistranslation | Other
> "Rotten little   bounder   too!" he said savagely.

> (acc)   Mistranslation | Other
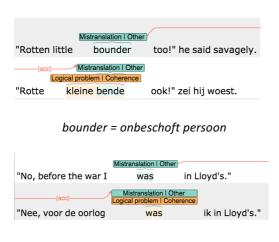>   Logical problem | Coherence
> "Rotte   kleine bende   ook!" zei hij woest.

bounder = onbeschoft persoon

> Mistranslation | Other
> "No, before the war I   was   in Lloyd's."

> (acc)   Mistranslation | Other
>   Logical problem | Coherence
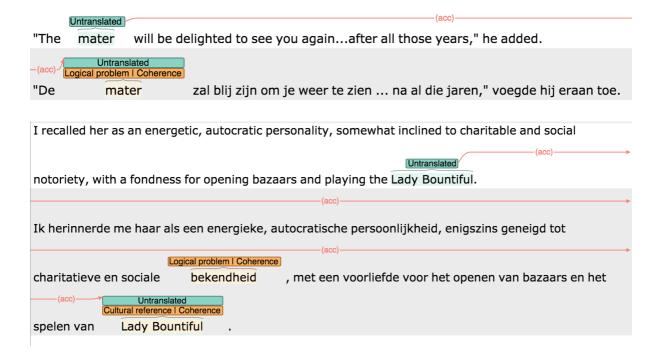> "Nee, voor de oorlog   was   ik in Lloyd's."

was = werkte … bij

Do Not Translate

The source content is unnecessarily translated into target language when it should have been left untranslated.

Untranslated

The source text is not translated (but was copied to target) when it should have been translated into Dutch.

> Untranslated   (acc)
> "The   mater   will be delighted to see you again…after all those years," he added.

> (acc)   Untranslated
>   Logical problem | Coherence
> "De   mater   zal blij zijn om je weer te zien … na al die jaren," voegde hij eraan toe.

I recalled her as an energetic, autocratic personality, somewhat inclined to charitable and social

> Untranslated   (acc)
> notoriety, with a fondness for opening bazaars and playing the Lady Bountiful.

> (acc)
> Ik herinnerde me haar als een energieke, autocratische persoonlijkheid, enigszins geneigd tot

> (acc)   Logical problem | Coherence
> charitatieve en sociale   bekendheid   , met een voorliefde voor het openen van bazaars en het

> (acc)   Untranslated
>   Cultural reference | Coherence
> spelen van   Lady Bountiful   .

Addition

Target text that is not present in the source sentence.

Wish you hadn't later."

Addition | Logical problem | Coherence

Ik wenste dat je het later niet had gedaan. '

Omission

Source content that cannot be found in target sentence.

Mistranslation | Semantically Unrelated | Omission (acc)

"Well, I've always had a secret hankering to be a detective!"

(acc) Mistranslation | Semantically Unrelated
Logical problem | Coherence

"Nou, ik heb altijd een geheime hang naar een detective!"

Capitalization & Punctuation

Error related to capitalization or punctuation.