
Translating the Present Perfect Continuous

Author: Miranda 't Hoen
Bachelor CKI, UU
3897745

Supervisors: Henriette de Swart
Bert le Bruyn
Martijn van der Klis

Final version
7,5 ECTS
May 31st, 2017

Utrecht University

Overview

1. Introduction 3.
    1.1 Time in Translation 3.
    1.2 Relevance in the field of AI 3.
2. The problem 3.
    2.1 The Present Perfect 4.
    2.2 The Present Perfect Continuous 4.
    2.3 Difficulties in translating 5.
    2.4 Expectations 6.
3. Method 6.
    3.1 General outline 7.
    3.2 Corpus description 7.
    3.3 Time Align 7.
    3.4 Categorization 8.
    3.5 Visualization 9.
4. Results 11.
    4.1 Quantitative analysis 11.
    4.2 Qualitative analysis 16.
5. Discussion 18.
    5.1 Of the results 18.
    5.2 Of the method 18.
6. Conclusion 19.
7. References 20.

1. Introduction

1.1 Time in Translation

For this bachelor thesis I joined Henriette de Swart, Bert Le Bruyn and Martijn van der Klis in their existing Time in Translation research. Their aim within this research is to find any rhyme or reason in the way the Perfect tense in the English language is translated into other languages. Their hope is to eventually be able to define a general meaning of the Perfect tense (see Time in Translation website). Using fragments of text from a large parallel corpus and algorithms made by Martijn van der Klis, they look at and label the ways in which the Present Perfect tense is translated into or from Dutch, French, German and Spanish.

In this thesis I will present a corpus analysis intended to provide additional data for the Time in Translation research, specifically about the translation of the Present Perfect Continuous. This specific version of the Perfect tense is a form with seemingly no equivalents in most other languages, leaving translators to come up with creative solutions to find a translation with a high translational equivalence to the Present Perfect Continuous. By analyzing these solutions in a large parallel corpus, hopefully, I will be able to find an answer to my research question, which says: which approaches are used most and which work best when it comes to the translation of the Present Perfect Continuous? And the sub question; how do these approaches differ between the Dutch, French, German, Spanish and Portuguese languages?

1.2 Relevance in the field of AI

The world we live in, is the most connected it has ever been. The internet enables all of us to communicate with anyone, including people on the opposite side of the world. The fact that this connectedness is a thing of the recent decades means that the first languages have initially developed with little to no communication between them, causing the structures of these languages to be vastly different. Since then, many more languages have come into existence, some of them more similar to each other than others, but still with fundamental differences. To be able to translate these languages to and from each other with as little loss of meaning as possible, we need a great understanding of both the meaning of a sentence in the source language, and of the target language.

As an AI student, this research sparks my interest for its many possibilities for innovation in the field of machine translation. Google Translate, the best known example of a machine translator, is an exceptionally helpful instrument to many (human) translators, language students, and a certain AI student writing a bachelor thesis about time in translations. In order for us to be able to improve the way these translators work, we need knowledge about the way human beings, preferably native speakers, would translate certain fragments of text. Native speakers are the most experienced in their language, and will therefore be able to construct translations with the highest level of translational equivalence, making their translations the examples we need. Developing machines that are able to simulate the way in which these native speakers translate sentences that have no one-to-one translation seems like a very exciting challenge and an ideal goal for the above mentioned users.

2. The Problem

In order to give an idea of the problem at the root of this research, I will first give an explanation of the Present Perfect, and the Present Perfect Continuous in section 2.1 and 2.2, after which I will point out the difficulties in translating the Present Perfect Continuous in section 2.3, and after that, in section 2.4, give a brief explanation of the expectations about the outcome of this research.

2.1 The Present Perfect

The Present Perfect tense in English, composed of "have" or "has" and a Past Participle (i.e. "Have worked"), is used to describe actions that occurred in the past and whose results are relevant in the present. There are multiple variations in the reading of sentences containing a Present Perfect.

1.a. I have worked at that company. (but I don't anymore)

1.b. I have promoted myself. (and now have a better job)

1.c. I have worked on this project for weeks. (and am still working on it)

Sentence 1.a. is the version of the Present Perfect whose meaning seems the most straightforward. It says that I started working at that company at some point in the past, but am not currently working there anymore. This reading of the Present Perfect has been labeled as an *experiential/existential* reading by Nishiyama and Koenig (Nishiyama and Koenig, 2010), making the 'working at that company' the experience that completely exists in the past, and that we can refer to in the present.

Sentence 1.b. describes an action that has started and ended in the past but has a consequence that is relevant in the present. This type of sentence is labeled by Nishiyama and Koenig as a *resultative* Perfect reading. The action that fully took place in the past has resulted into a new state of being (having a better job) in the present. This currently relevant result of the past action is what sets this type of reading apart from the experiential one, in which there is no current result of the action, merely a reference to it.

The Present Perfect can also be used to describe continuous actions that started in the past, are still occurring in the present, and will most likely continue to occur in the future. Sentence 1.c. is an example of this. The work on the project has started weeks ago, is still being worked on at the time of the utterance and will need work in the future. In order for the continuous meaning to be expressed in a sentence with a Present Perfect, an indication of time will have to be added. In the example above "for weeks" is an indication of time, as is an indication like "since yesterday", for example. Nishiyama and Koenig labeled this reading of the Perfect *continuative*.

2.2. The Present Perfect Continuous

The English language also contains a Present Perfect Continuous tense which, from now on, I will refer to as the PPC. The PPC can be seen as a specific construction for references to continuous actions, like in sentence 1.c in the previous section. The PPC is formed by the Present Perfect form of 'to be' ("have been" or "has been") followed by a Present Participle, which is a participle that ends in '-ing'. An example of a PPC would be "I have been working on this project". This sentence inherently means that I have started working on this project at some point in the past, and that I am continuing this work in the present. If one wants to give information about *when* the work on this project has been started, they will again have to add an indication of time, since the PPC on its own can only refer to an unspecified moment in the past. As we will later see, this indication of time will return as a point of significance within this research.

The PPC as described here can only be formed with the Present Participle form of dynamic verbs, ending in '-ing', which describe temporary *actions* that can change over time (i.e. "He has been walking"). The PPC can therefore not be formed with so called stative verbs, which are verbs that describe *states* of being, of which its '-ing' form would not have the correct continuous meaning when paired with a Present Perfect. If one were to try to form a PPC using a stative verb, they would get an ungrammatical sentence like "he has been knowing". This sentence sounds odd since knowing something is a permanent *state;* once you know something you cannot un-know it.

There are two types of stative verbs; stage-level and individual-level, and even though neither of them are suited to be used in the original PPC, they can be used in a different way to express continuity. Stage-level verbs are used to describe temporary states, whereas individual-level verbs are used to describe permanent states. The unsuitability of these types of verbs in the form of a Present Participle can be seen below:

Stage-level:	✓ 'I don't mind'	× 'I am not minding'
Individual-level:	✓'I didn't know'	× 'I wasn't knowing'

To construct a continuous sentence using stage-level verbs is challenging, since these verbs are mainly used in experiential readings of the Perfect (Nishiyama and Koenig, 2010), which means they are generally used to describe actions that have started and ended in the past – not continuous actions. Because the individual-level verbs describe permanent states however, their use in a Present Perfect would automatically result in a continuous reading (i.e. "I have known (him for years)" (and still do)).

Initially, the plan for this research was to extract and analyze the translations of both the PPC and the individual-level Continuous, to see whether there would be any differences in the way these occurrences had been translated. Due to a lack of time, the decision was made to solely focus on the translations of the PPC in this corpus analysis, leaving the individual-level Continuous for a future (extension of this) research.

2.3. Difficulties in translating

Anyone who has learned English as a second language, or any native speakers who have ever gone through the process of learning another language might recognize the Present Perfect (Continuous) as a big discrepancy between English and most other languages. The apparent absence of the PPC's form in most other languages makes it difficult to grasp the PPC's meaning, let alone translate it into a (seemingly less expressive) target language. I will use my native language in an example. If I were to try to translate the PPC-phrase "I have been driving" into Dutch, my options would be:

1. Ik rijd. (= *I drive*)
2. Ik ben aan het rijden. (= *I am driving*)
3. Ik reed. (= *I drove*)
4. Ik heb gereden. (= *I have driven*)

None of these translations are capable of expressing the continuity of the PPC, neither in Dutch, nor in English. Sentence 1 and 2 are missing the Past aspect of the PPC-phrase, and sentence 3 and 4 are missing the Present aspect of the PPC-phrase. The two Present-sentences, 1 and 2, and Past-sentence 4 could however adequately express the PPC's continuity if combined with a time indication, as seen in sentence 1.c. in section 2.1.

1. Ik rijd <u>al uren.</u> (= *I drive for hours already*)
2. Ik ben <u>al uren</u> aan het rijden. (= *I am driving for hours already*)
4. Ik heb <u>al uren</u> gereden. (=*I have driven for hours already*)

Even though the English literal translations of these sentences either do not quite seem to capture the continuity of the PPC or just sound odd, I can tell you that in Dutch they are commonly used translation equivalences of the phrase "I have been driving for hours already". Apart from being an example of why these time indications are so important in these types of translations, this also shows us why native speakers, or at least people proficient in the relevant languages, are needed in these types of research.

In order to be able to successfully translate a PPC, where no one-to-one translation is available, a person needs context, creativity and a great level of understanding of both the meaning of the PPC in question and the target language. One will need to find a way to express the continuous meaning of the PPC using the available tenses in the target language. In this thesis we will take a look at the methods that translators use to preserve the meaning of the PPC as best as they can, and whether these methods suffice.

2.4. Expectations

In the first original Time in Translation analysis of the Perfect in EuroParl, three instances of the PPC were included in the results, which is what one of the expectations for this research is based on. All three occurrences, of which one can be seen in figure 1, were translations with German as the source language and English as one of the target languages. The time indications underlined in purple combined with the highlighted green words in the Dutch, French and German fragments give those sentences the continuous meaning that is expressed by the PPC in the English and Spanish fragments, as explained in the previous section.

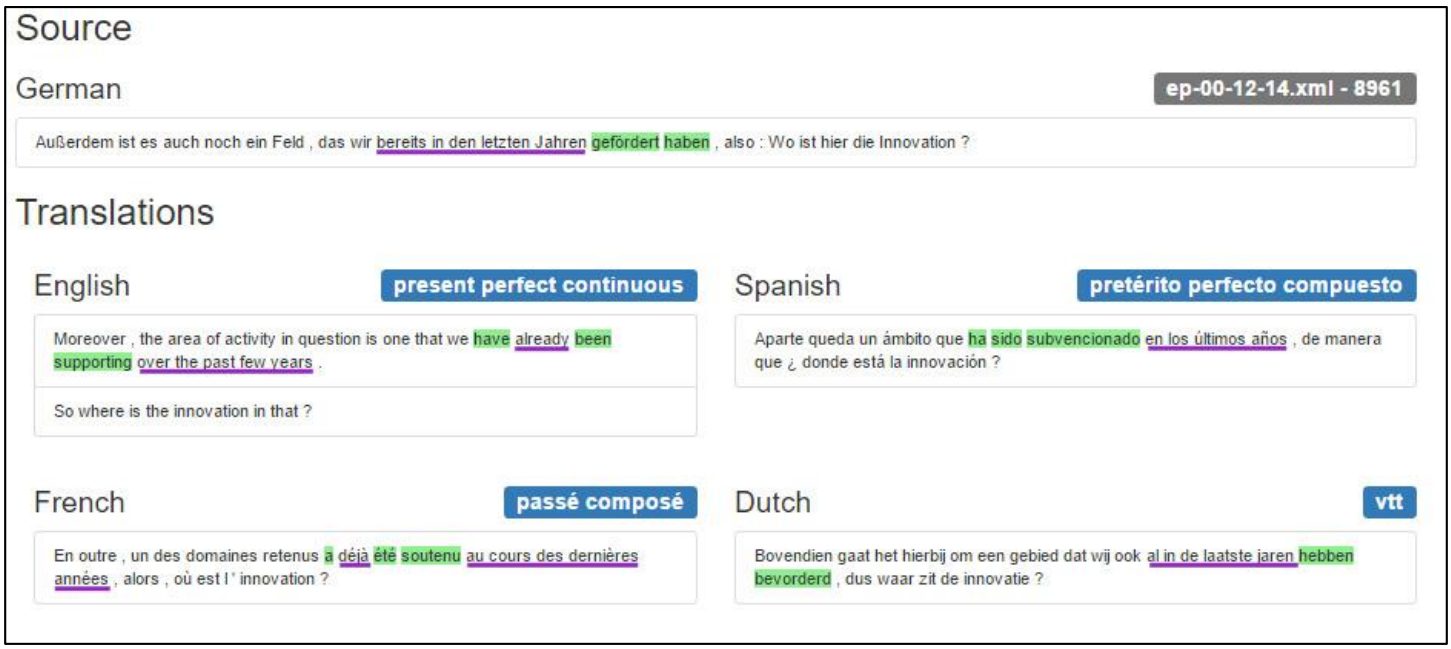


Figure 1 One of the three occurrences of the PPC in the original Time in Translation research.

Taking note of this, it is fairly possible that for the English translator, the time indication functions as kind of a trigger to use the PPC, even though the time indication itself is not omitted in the English translation. In the cases in which the English fragment contains a PPC without a time indication, the parallel fragments are expected to have a time indication added to whichever tense they used to translate the PPC with, in order to preserve the continuity within the translation.

Besides the use of time indications, one reasonably expects to find context-dependent creativity within the use of tenses and perhaps even verbs. Nominalization is an example of the latter, where a verb is used as a noun, making this one solution that is expected to be found in translations in which the continuity of the PPC is not necessary to the meaning of the sentence. The phrase “the video we have been watching” could be translated to the target language’s equivalent of “the watched video”. Apart from this, as far as creativity goes, any actual predictions about the translations or differences between the languages would be unsubstantiated.

3. Method

In this section, a description will be given of the methods and tools used within this research. In section 3.1 I will describe the general outline of the taken approach, section 3.2 will provide a brief description of the used corpus, section 3.3 will provide a visual guide of the Time Align application, section 3.4 will describe the process of categorization and finally, in section 3.5, I will describe the process of visualizing the data.

3.1. General outline

The general setup of this research is the same as the setup of the initial Perfect analysis in the Time in Translation project. The same corpus is used and a slightly modified version of the same algorithm is used to extract fragments from it. This research is different in the tense on which the fragments were selected – for this research the algorithm was configured to find occurrences of the Present Perfect Continuous, instead of Simple Present Perfects. Because the form of this tense seems to be mostly exclusive to the English language, the source fragments will solely be extracted from the English texts. In the original research, occurrences of the Perfect in five languages were analyzed, which in this research would be reduced to four since the English fragments will only serve as source fragments. In order to keep the amount of accumulated data roughly the same, the decision was made to include another language. Portuguese was chosen out of a personal interest and average proficiency of mine.

3.2. Corpus description

The corpus used for this analysis is EuroParl (see website EuroParl corpus in references), which is a parallel multilingual corpus containing speeches from meetings at the European Parliament and their many translations. This corpus was chosen for its vastness and reliability, since these translations are executed by professional translators and rigorously checked. Using the Perfect Extractor, about 250 English fragments containing a PPC and their German, Dutch, French, Spanish and Portuguese translations were extracted from a quarter of a year of speeches. (Specifically: Q4 of 2000, which is often used for this type of research.)

3.3. Time Align

Time Align is an application made by Martijn van der Klis, enabling a user to annotate extracted parallel fragments of text. The application can be utilized on the website of the Time in Translation project (see website in references), under the heading 'translation mining'. Once the corpus texts have been uploaded and the user has opened the Time Align application, the screen will look like in figure 2. On the left (1) the user will be presented with the original fragment, which in the case of this research is always English. On the right (2) is the parallel translation of this fragment in the language selected by the user. The automatically highlighted green words in the original fragment form the PPC by which the Perfect Extractor has chosen to extract this fragment, and the highlighted green words on the right side are manually selected by the user as the translation of the PPC on the left.

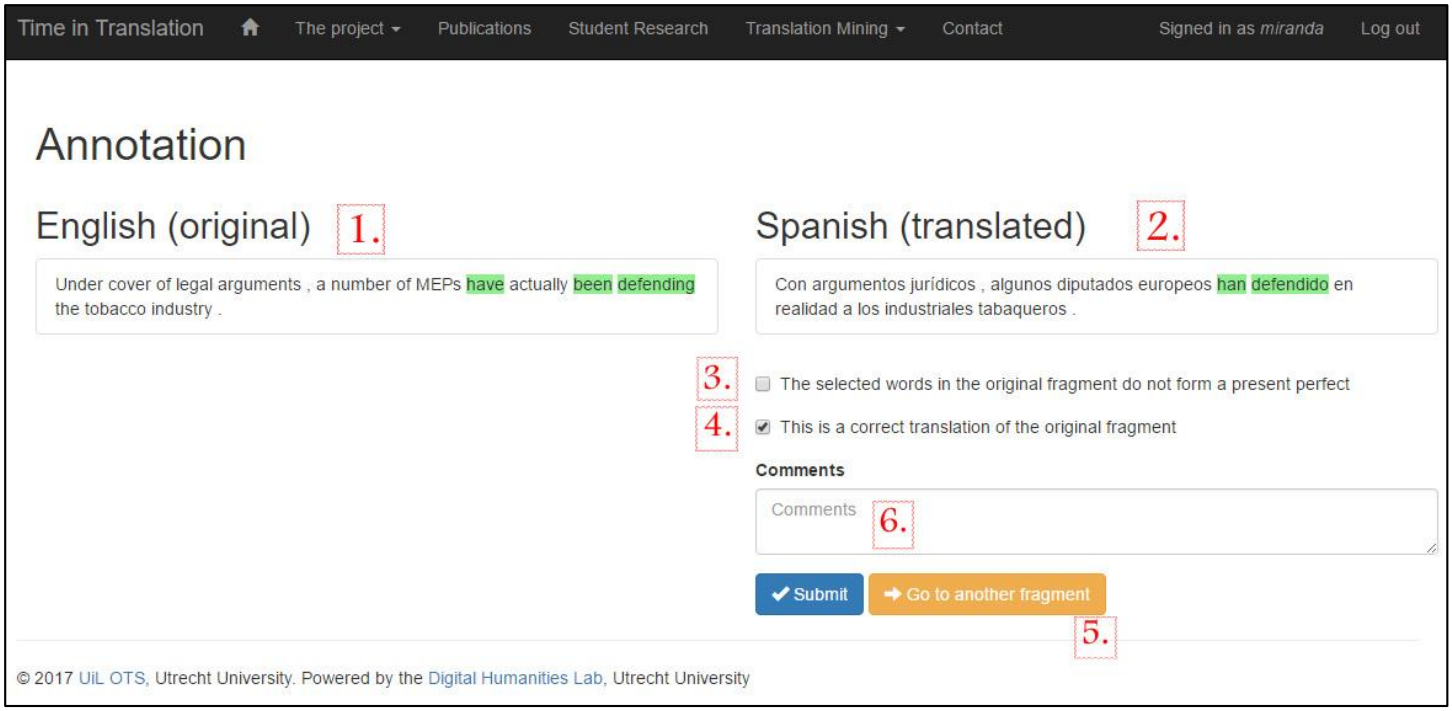


Figure 2. The Time Align application in use for an English-Spanish translation on the website of the Time in Translation project.

If the highlighted words in the source fragments represent a false positive, and thus do not represent a PPC, the user can report this by checking the first box (3) underneath the target fragment, which says "The selected words in the original fragment do not form a Present Perfect". In this research, this box has been checked particularly often, because of the way the PPC was 'recognized' in the English fragments. The Perfect Extractor looked for phrases of the form "have been" or "has been" followed by a verb ending in -ing. Even though this worked well (many PPCs were found), it also resulted in the application making selections like the one below in figure 3, which does not qualify as a PPC.

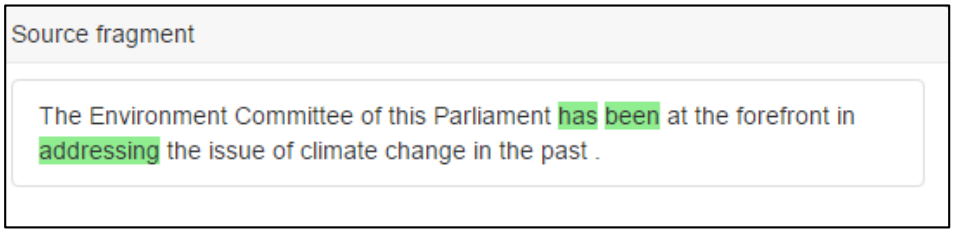


Figure 3. An example of a non-PPC, selected as a PPC by the Perfect Extractor

For all correctly selected PPC's, once the selection of its translation in the target fragment has been made, these words can be submitted as the correct translation of the presented PPC. If the fragment on the right is not a correct translation of the source fragment, however, this can be declared by unchecking the second box (4) underneath the target fragment, which says "This is a correct translation of the original fragment".

If the user wishes to skip the current fragment to be re-evaluated at a later time, he or she can click the "Go to another fragment"-button (5) instead of "Submit". The user is also able to add a comment to an annotation in the comment box (6).

3.4 Categorization

After the process of annotating in the Time Align application, the verb phrases that were selected as translations of the PPC's in the English fragments are organized into an excel-file and categorized. Initially, four categories were chosen but due to the creative nature of the translations, many more were added during the process. However, because not every category was present after a selection of result-relevant translations was made, some categories were dropped again as well. The translations were eventually grouped into the following categories:

Past, Present, PastPerf, PresPerf, RecentPast (only FR), Gerund (only SP & PT) and Cont (only SP & PT). Examples of translations in every category can be seen in figure 4.

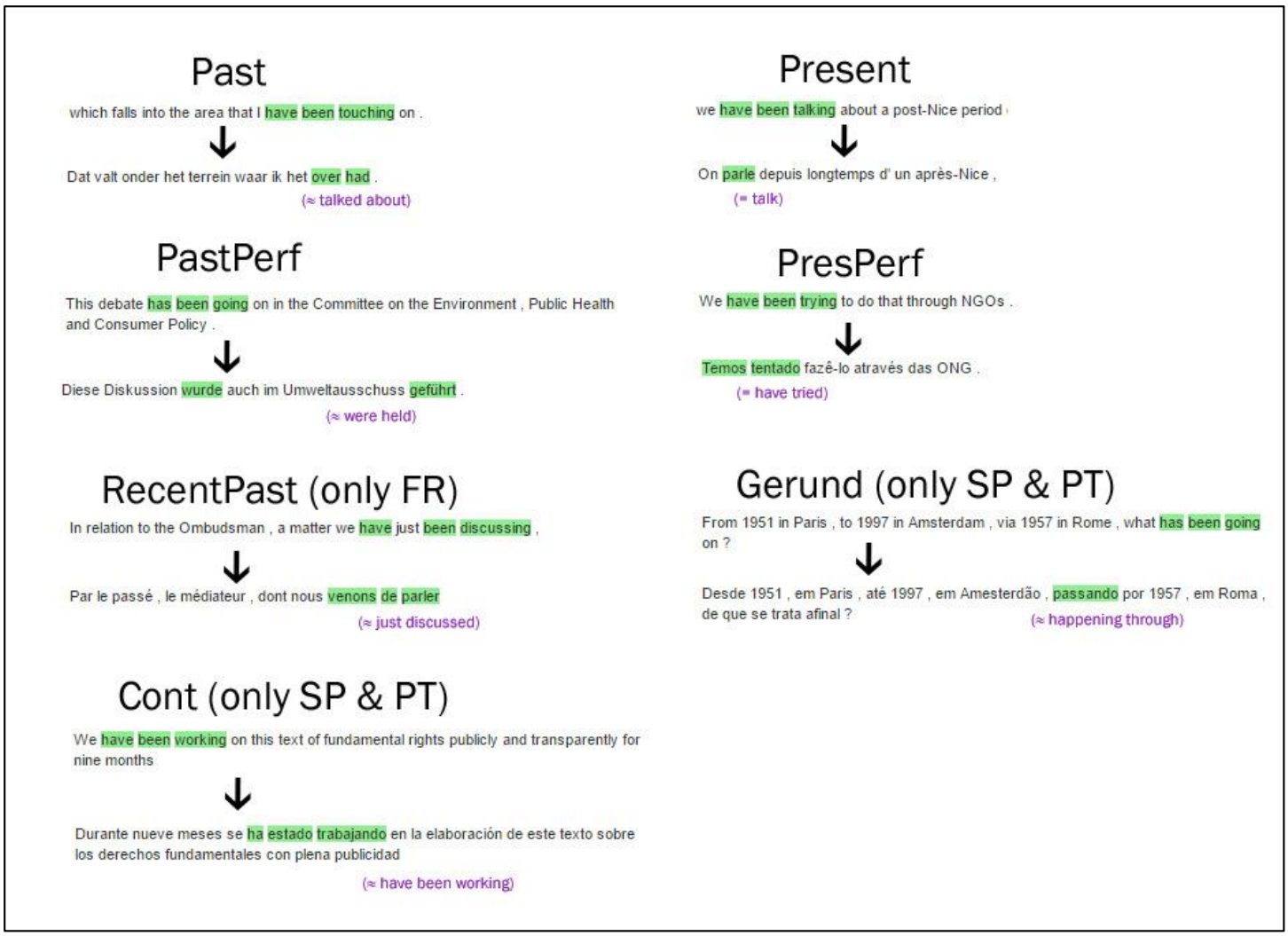


Figure 4. An overview of the categories used in this research, each with an example translation underneath it. The purple text is a (rough) translation of the highlighted text in the target language.

3.5 Visualization

The translations are represented as six-tuples of categories, with every language having its own position in the tuple; <French, English, Dutch, Portuguese, German, Spanish>. Because the decision was made to use only the tuples containing translations for all six positions in the visualization of the results, a lot of tuples had to be dropped, since a lot of the initial tuples had missing (selected as incorrect) translations for one or multiple languages. An example of what the remaining tuples look like can be seen below in figure 5.

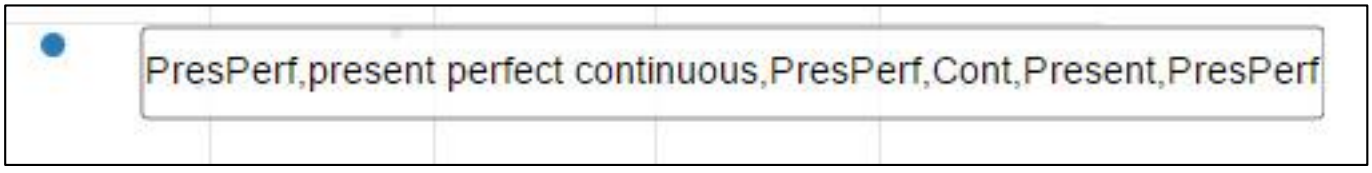


Figure 5. A six-tuple of categories.

Using the same distance function used in the initial Perfect-research (van der Klis, Le Bruyn and de Swart, 2017), tuples are assigned a distance value based on the number of mismatches in categories between the tuples. If all six categories in two tuples were to be the same, for example, the distance value would be 0. For every mismatch between the two tuples, 1 is added and the total divided by 6 is the value assigned to it. Once all tuples are given distance values they can be used to construct a dissimilarity matrix. A visualization of this matrix can be made using

Multidimensional Scaling (Wälchli and Cysouw, 2012), which will result in a scatter plot like the one below in figure 6, which shows results from the original Time in Translation research. The results of this research will be presented in the same way in the next section.

Figure 6. A scatter plot of the results for the German language in the initial Time in Translation research, visualized with Multidimensional Scaling.

The dots in the scatter plot (1) represent the data points, with their colors offering a helping hand in easily visualizing the clustering within the plot. Each color represents a tense, as can be seen in the legend (2), which stay the same for these tenses in the other languages, making the plots easier to compare. The colors in the legend can be clicked, toggling the appearance of that particular tense on or off. At (3) one can select the language to be plotted, and the dimensions of the plot can be adjusted at the bottom (4) as well. Once the selection of language and dimensions has been made, the user clicks the green button and the new plot will appear.

When the user hovers over a data point in the plot, the corresponding tuple appears, as in figure 5. The user can also click on a data point to be brought to the fragment overview where they can see the tuple in its entirety, with all selected phrases in the original fragments, in every language. This looks like figure 7.

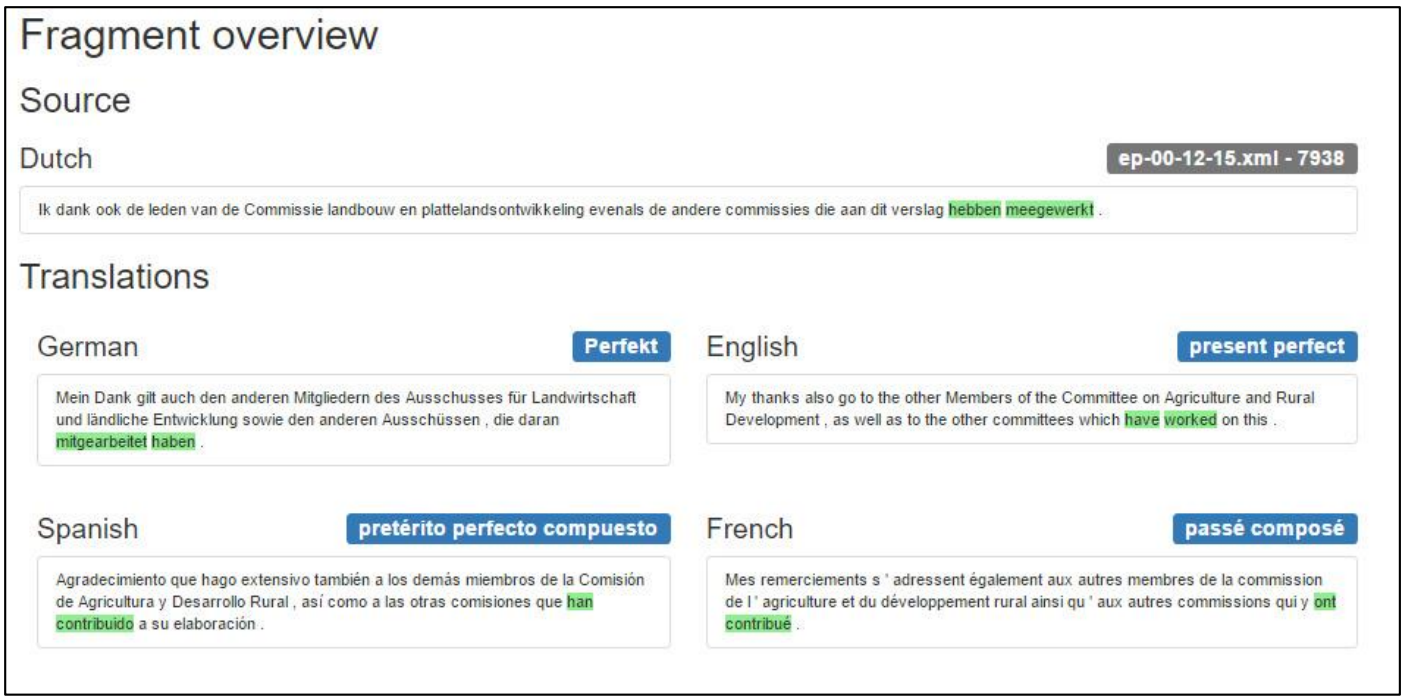


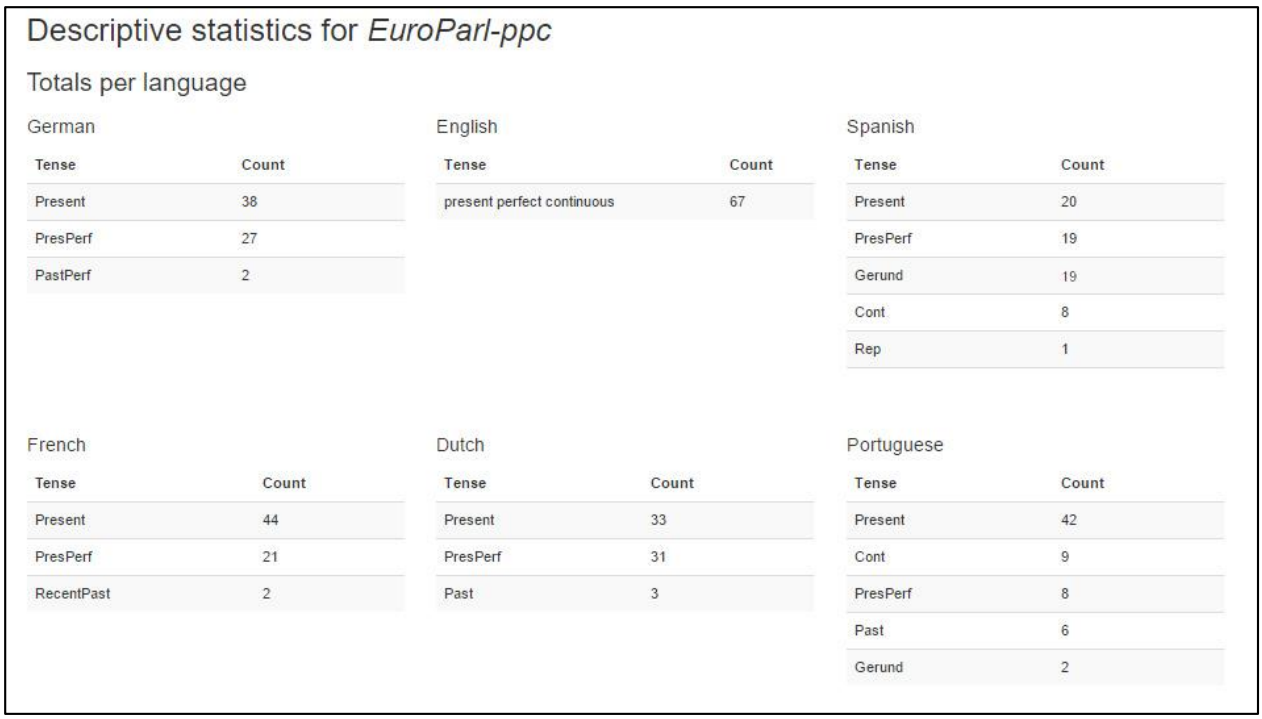
Figure 7. A fragment overview of a single data point in the scatter plot, showing the five-tuple in its entirety, including all fragments, annotations, categories, and indication of source fragment.

4. Results

In this section, the results of this research will be presented. Section 4.1 will show the results of a quantitative analysis, followed by the results of a qualitative analysis in section 4.2. Section 4.2 will also contain some initial observations which were made during the phase of annotating fragments and during the process of categorizing.

4.1 Quantitative analysis

Figure 8 below shows the descriptive statistics about the categorized translations. This overview can be found in the Results-section on the Time in Translation website, and shows us how many translations every category in every language contains. As can be seen in the middle, only 67 fragments of the initial 250 fragments created tuples in which all five languages provided an adequate translation of the corresponding PPC, giving us 67 data points to construct a dissimilarity matrix with.



Descriptive statistics for *EuroParl-ppc*

Totals per language

German

Tense	Count
Present	38
PresPerf	27
PastPerf	2

English

Tense	Count
present perfect continuous	67

Spanish

Tense	Count
Present	20
PresPerf	19
Gerund	19
Cont	8
Rep	1

French

Tense	Count
Present	44
PresPerf	21
RecentPast	2

Dutch

Tense	Count
Present	33
PresPerf	31
Past	3

Portuguese

Tense	Count
Present	42
Cont	9
PresPerf	8
Past	6
Gerund	2

Figure 8. Descriptive statistics for the data in this research. This overview can be found in the results-section on the Time in Translation website.

It is interesting to see how in every language the Present tense was used most often to translate a PPC with, followed closely by a Present Perfect in all languages except Portuguese. Portuguese seems to be the outlier in most categories, judging by this schema. It is the only language where the Present tense is used more than four times as often as the next-in-line tense, all other languages seem to have a better 'spread'.

German, French and Dutch have very similar tables, with similar distribution over the three categories and only an interesting difference in the least used tense, which is three different variations of a Past tense. Spanish and Portuguese being the two languages with five remaining categories instead of three have their own similarities in distribution, the only difference being the usage of the Gerund which seems to be way more prominent in the Spanish translations, causing a shift in distribution over the Present- and Gerund-categories between these two languages.

The results will be presented language by language, showing the mapped data visualized with the use of Multidimensional Scaling executed on the dissimilarity matrix that followed from the categorizing of phrases in that particular language. For every language, the scatter plot will be shown in the same dimensions in order to be able to accurately compare the data between languages.

French

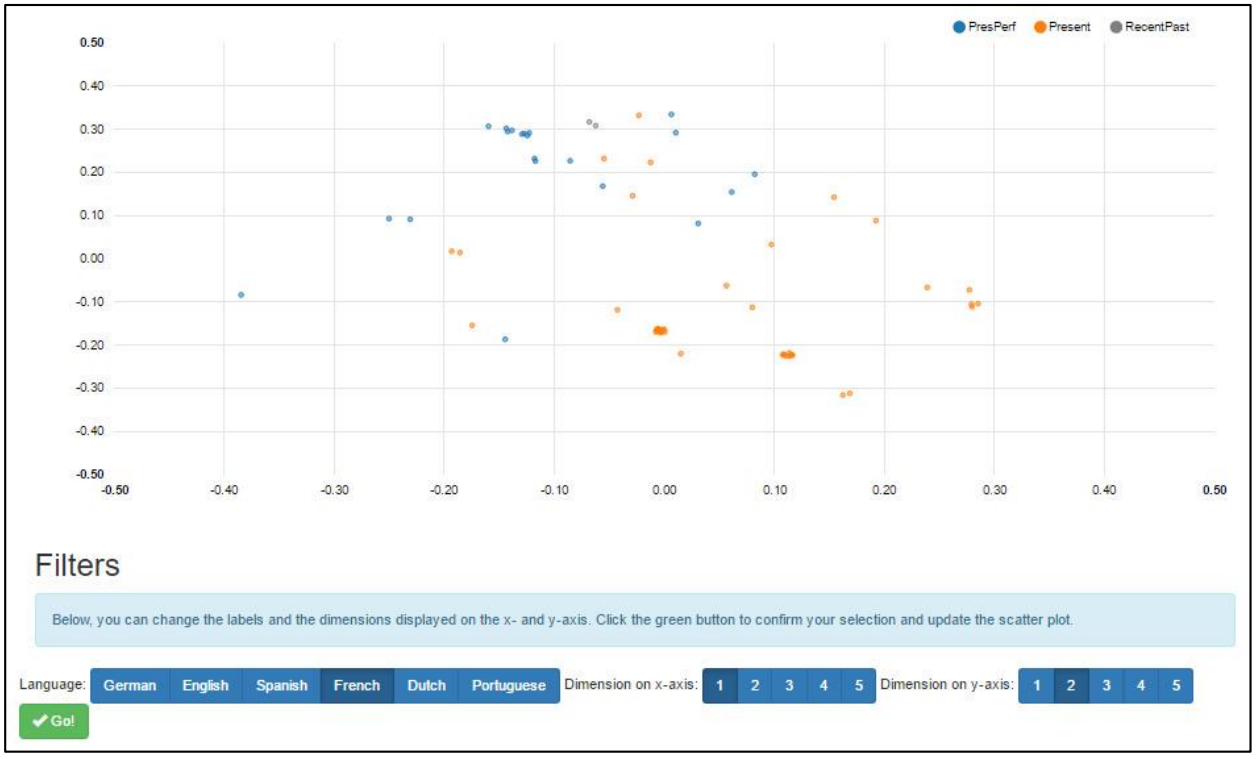


Figure 9. The scatter plot presenting the results of the French dataset.

Dutch

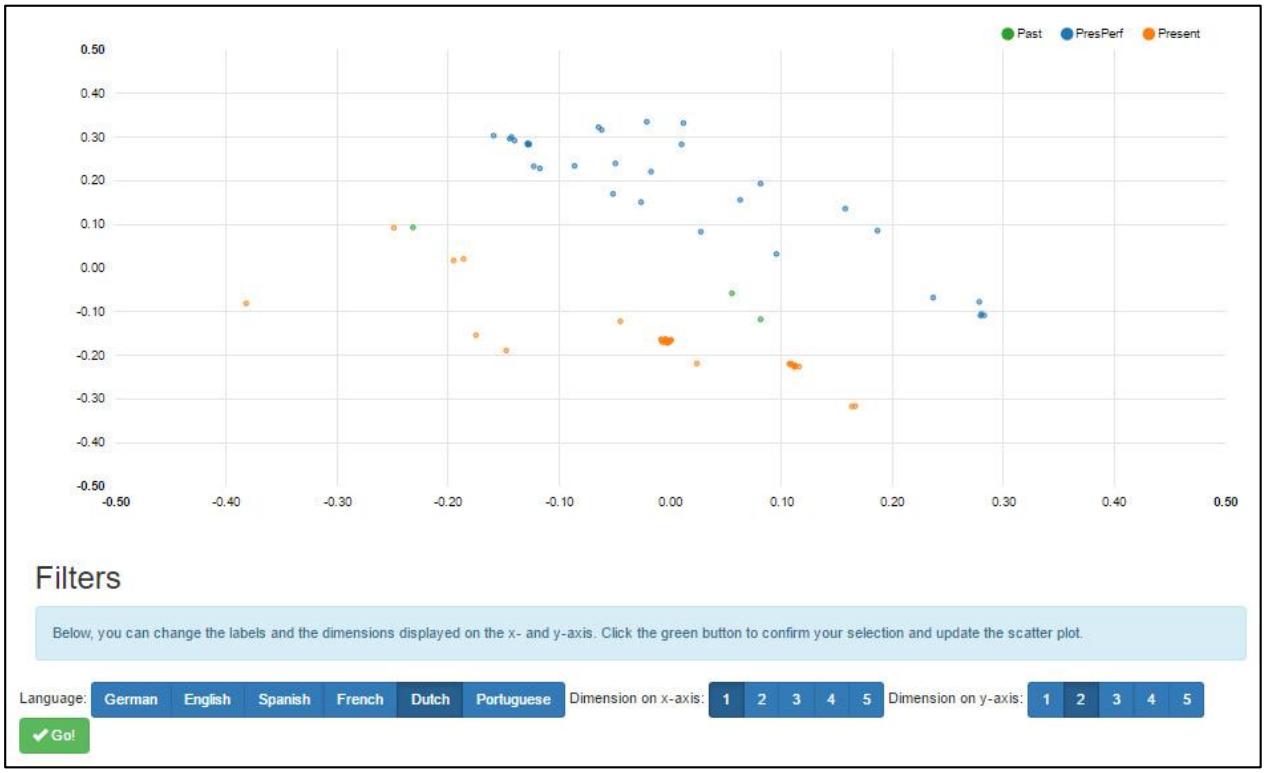


Figure 10. The scatter plot presenting the results of the Dutch dataset.

German

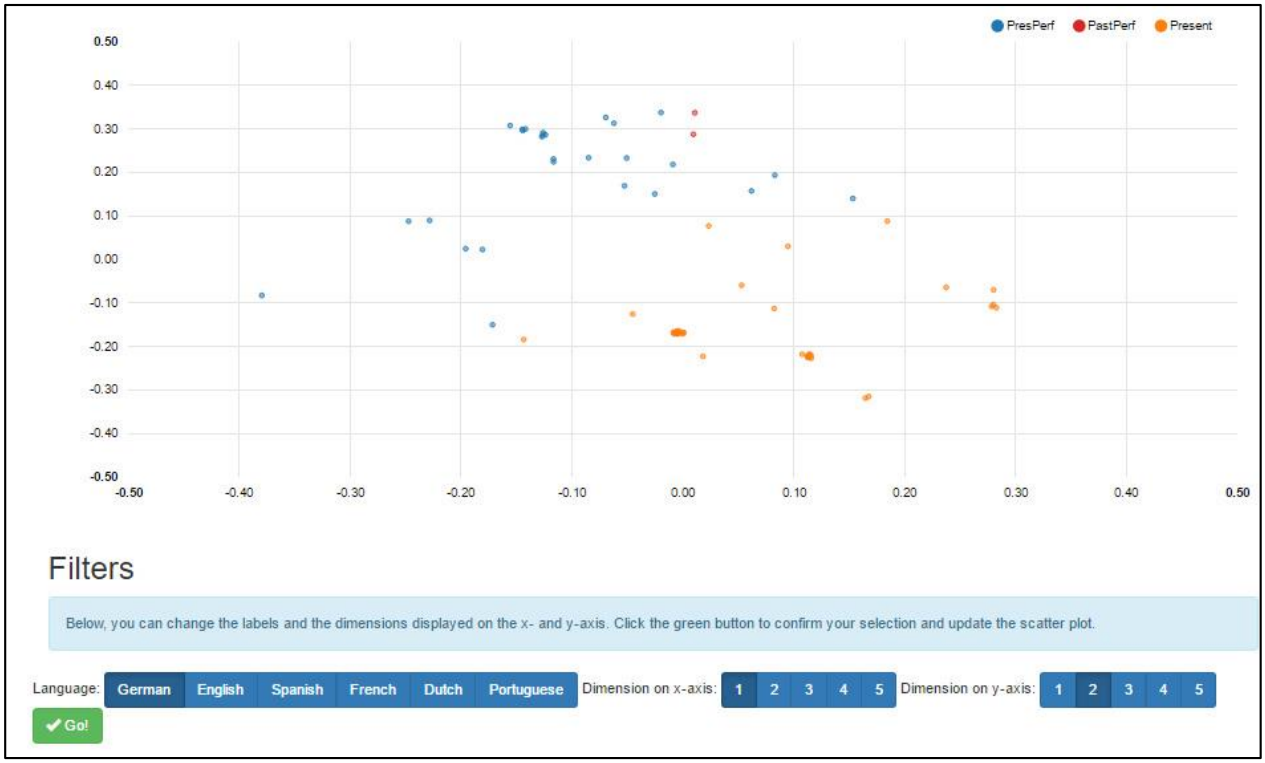


Figure 11. The scatter plot presenting the results of the German dataset.

Spanish

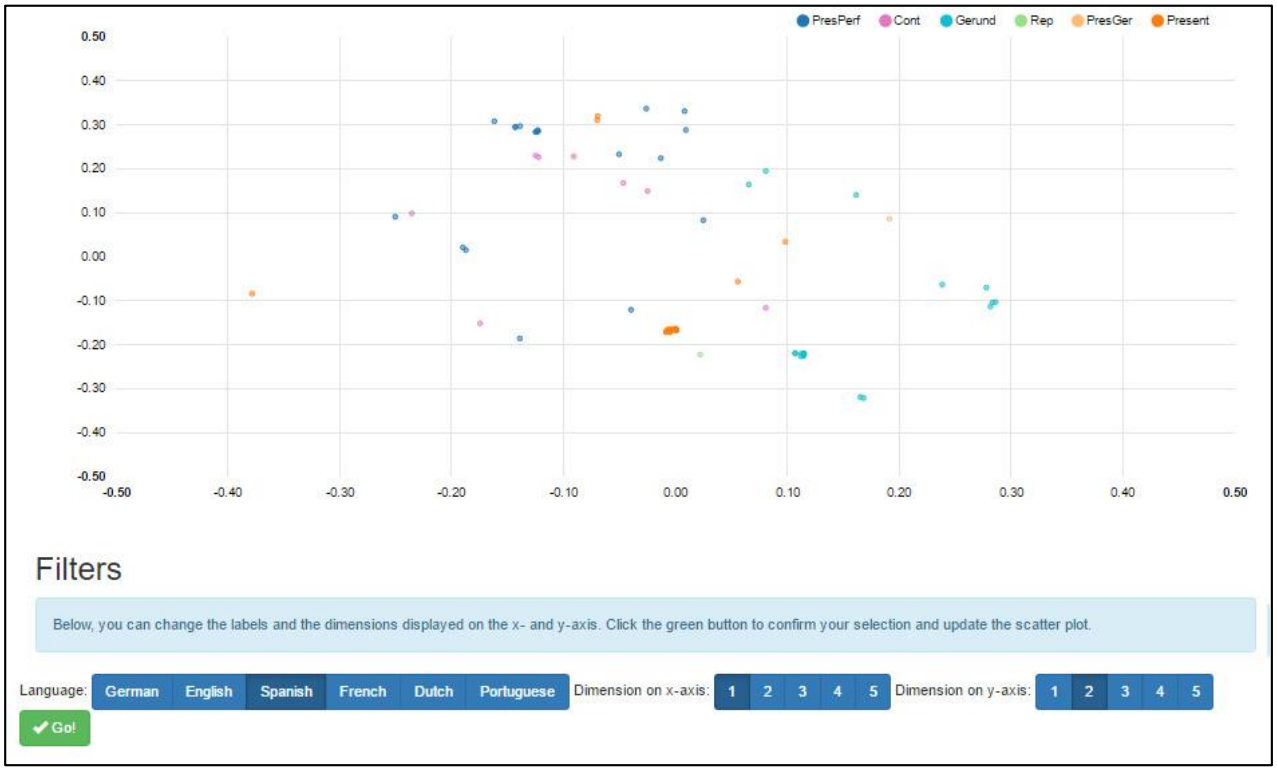


Figure 12. The scatter plot presenting the results of the Spanish dataset.

Portuguese

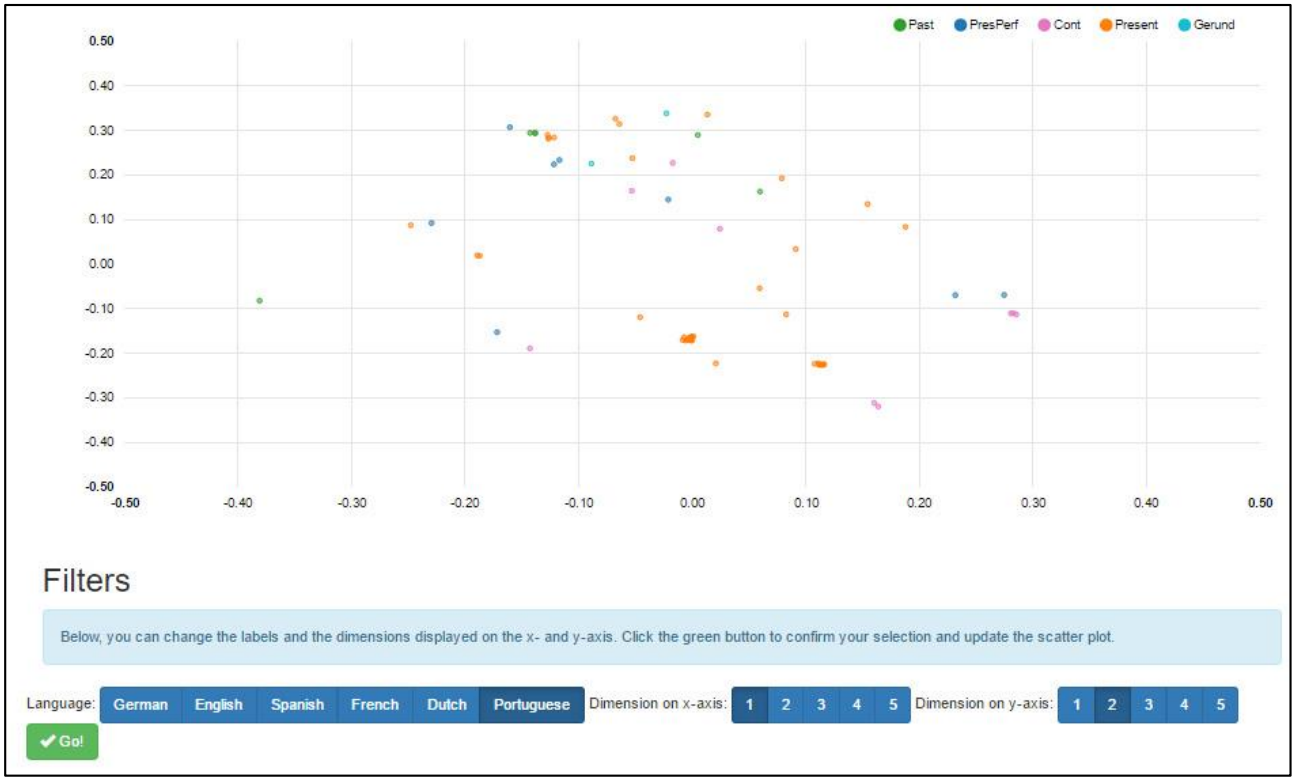


Figure 13. The scatter plot presenting the results of the Portuguese dataset.

These scatter plots provide relevant information mostly in the clustering and placement of these clusters. Quoting Henriette de Swart, Bert le Bruyn and Martijn van der Klis on the use of Multidimensional Scaling; *"This visualization shows which space the various tenses (PERFECT and other) occupy on the map, and thus enables researchers to see how tenses interact within a language."* (van der Klis, Le Bruyn and de Swart, 2017)

These plots also provide us with a way to visually compare the behavior of all relevant languages. The plots shown above tell us that the languages in this research all behave in a similar way, with especially the ones for the French and German data being nearly identical. Both have the Present tense clustered in the bottom right corner and the Present Perfect tense clustered in the upper left corner, with the French data having some more stray Present occurrences in the Present Perfect section. The Dutch data seems to represent a mirrored version of the French and German plots. To see this more clearly, I added some rough division lines in figure 14.

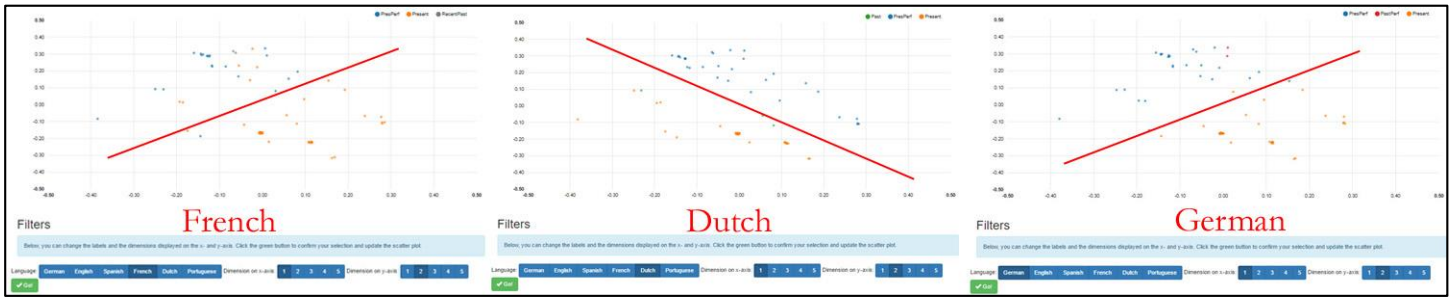


Figure 14. The scatter plots of the French, Dutch and German data with red lines roughly indicating the division between clusters.

The Spanish and Portuguese plots also seem very similar, but this is harder to see because five categories, and therefore five colors, are represented in them. In order to make the comparison with the French, Dutch and German data a little easier, I momentarily switched off the visualization of the three extra categories.

The plots in figure 15 below show the distribution of solely the Present tense and the Present Perfect tense in Spanish and Portuguese, for convenience immediately including the rough division lines:

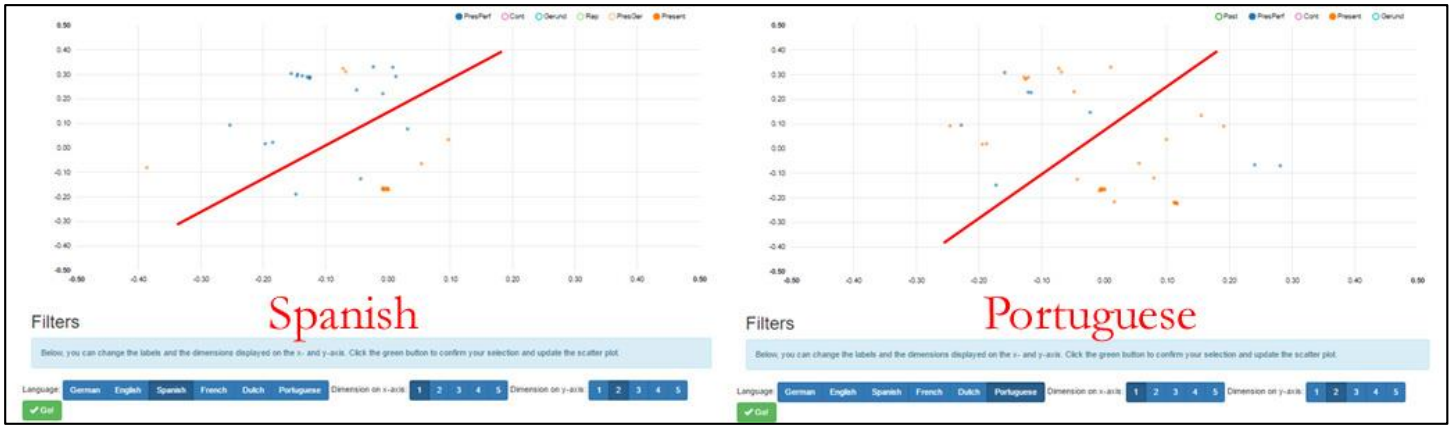


Figure 15. The scatter plots for the Spanish and Portuguese data, with division lines.

4.2 Qualitative analysis

Apart from having these maps and statistics to provide quantitative information about the translation of the PPC, some qualitative observations were made during the process of annotating and categorizing which might also be able to provide valuable insight into the translation of the PPC. The seemingly most relevant ones will be mentioned in this section.

The first thing that was noticed was the usage of time indications in all fragments; both the English source fragments and the fragments of all target languages. As explained in section 2.1, the Simple Present Perfect requires a time indication in order to express continuity. The PPC, however, does not necessarily need one. The amount of time indications in the English fragments used within this research is therefore peculiar, raising the suspicion that English was rarely ever the source language of the original speeches. Whenever continuity was expressed in the actual source language of the speech, inevitably using time indications due to a lack of PPC-form in that language, these time indications were also translated into English. They most likely also worked as a kind of 'trigger' for the English translator to use a PPC, resulting in a sentence containing both a PPC and a time indication.

This observation led to an eventual urge to label the English fragments with whatever time indication was present in the fragment (i.e. "for the past ten years"), or a "no" when a time indication was missing. 45 of the 67 English fragments were eventually labeled with a time indication.

This was done with the intention of analyzing the use of these time indications in the other languages, having hypothesized that any time indication present in a target fragment, despite there not being one in the English fragment, would provide us with relevant information about the translation. If an English fragment does not contain a time indication but the translation does, the time indication was most likely added to aid in the preservation of the meaning of the PPC.

Therefore, the occurrences of these 'rogue' time indications in the target languages would be an interesting point of research, since they seem to be the easiest and most frequently used way of translation a PPC to another language. Within this research, of the 22 English fragments without a time indication, none of the parallel fragments in the other languages contained a time indication either. A possible explanation for this will be given in the next section.

The second interesting observation was made during the process of annotating and revolves around just a single Spanish translation, which will be presented after some examples of the other creativity that was used. As can be seen in the descriptive statistics in FIGURE, all five target languages mainly used the Present tense to translate the PPC, of which the majority was in combination with the above mentioned time indications, as were the Past, PresPerf and PastPerf. The outliers, the fragments in which no time indication was present, are therefore the most interesting cases to look at, since this is where the translator's creativity will have come into play.

There are cases in which the translator seems to attempt to avoid the translation of the PPC, for example by getting rid of the verb phrase all together and expressing the original sentiment by transforming it into an adjective:

Figure 16. A fragment translated by expressing the PPC-phrase by using the verb as an adjective.
In which the German fragment roughly translates to "One of the mentioned sensitive areas is trade policy".

Or sometimes by replacing the PPC-verb with an entirely different verb:

Figure 17. A fragment translated by replacing the PPC-phrase with a different verb.
In which the Dutch translation of the PPC roughly translates to "address" or "refer to".

No creative translation seemed to be able to capture the continuity that the PPC expresses, though, they merely succeeded in translating the general subject of the phrase. One Spanish translator seems to have provided an adequate translation of the fragment in question, however. It was created using repetition, which was what its solitary category was called, as well.

Figure 18. A creative spanish translation of the PPC using repetition.

The expression "combatimos y luchamos" translates to "we fight and fight", which seems to be able to capture the continuity of the PPC "have been combating" beautifully. It tells us the speaker has fought, is fighting, and the phrase carries a sense of a future promise of this repetition. This translation was an exciting find, feeling completely surprised by an adequate translation of the PPC with near perfect preservation of the continuity, without the necessity of the time indication in the sentence.

5. Discussion

5.1. Of the results

The goal of this thesis was to find an answer to the research question posed in the introduction; which approaches are used most and which work best when it comes to the translation of the Present Perfect Continuous? and the sub question; how do these approaches differ between the Dutch, French, German, Spanish and Portuguese languages? In section 2.4 I hypothesized that the approaches would be of a creative nature, since the translators have to find a way to express the continuity of the PPC without being able to use an equivalent tense form.

One of the expectations was that they would make use of time indications to translate the continuity, because the addition of them can give a continuative meaning to a Present Perfect, as explained in section 2.1. Whether using time indications is a frequently used approach is difficult to conclude from the results of this research. Time indications were prevalent in the extracted fragments, but whenever they were present in a fragment of one of the five compared languages, the time indication was also present in the English fragment.

When looking at the 22 English fragments in which no time indication was present, however, no time indications were found in the parallel fragments either. This may be due to the possibility that the English fragments were never the source language of the speeches in these 22 cases, causing the absence of time indications in the source language to be translated into English. The data in this research is insufficient to draw any conclusions about the use of time indications, but I am of the opinion that it would be an interesting subject for future research.

Furthermore, it seemed that translators try to find a way around the PPC by using nominalization or by substituting the PPC with a different verb. These translations were not analyzed within this research and could, quite possibly, be another interesting subject for future research.

As far as differences between the languages go, all five analyzed languages in this research seem to behave rather similarly, with especially the French and German languages being nearly identical in their approaches. The division lines added in the scatter plots in section 4.1 created the impression that the Dutch language seemed to portray opposite behavior, with its division line mirrored to those of the other languages. I think the difference is not as big as it may seem, the Present Perfect cluster is still mainly in the upper half of the plot and the Present cluster still in the bottom half. The tilt of the division line is a result of the clusters having 'bled' into the other side (right/left).

Other differences include the Spanish and Portuguese having and using a continuous form to translate the PPC with, and these two languages using gerunds. The Spanish language was also the only one to showcase the use of repetition, which was the only creative solution that adequately preserved the continuity of the PPC.

5.2 Of the method

In this research the main problems arose within the belated realizations of a better or more convenient order of actions. When the time came to categorize the annotations, I wished I had thought of these categories before annotating, which would have eliminated the work of having to translate every sentence in the languages that I'm not great at, twice. Which leads me to a different point, which is that I am of the opinion that this kind of research is better to be executed by people with an extensive knowledge of all languages used, for I was having a lot of trouble finding out which tenses were being used in the translations. Unless it's possible to have the research executed by a person who is proficient in all relevant languages, it might be best to do the categorizations with the help of native speakers. I was lucky to have the help of my supervisors, who corrected some of my categorizations in the languages they know.

I also wish I had thought of the relevance of no-translation translations before. When annotating, I mercilessly threw out the fragments which did not contain a correct translation of the English fragment. However, this included all translations made with nominalization and other creative solutions, which I would, knowing what I know now, have put into categories of their own. Perhaps if I had been a little less radical in throwing out certain translations, the results of this research could have been based on more than just 67 fragments, making any conclusions drawn from the data a little more stable.

As great as EuroParl is as a corpus for this kind of research, due to its vastness and reliability, I do wish I had had access to information about the language in which the speeches were originally made. This kind of information would have provided me with insight into whether the English fragments were ever the source language, or if they were translations of another language in which the PPC was added. Perhaps this information is actually available in this corpus, I do highly recommend anyone working on a similar project to extract and use this information in order to make accurate observations.

Lastly, because I eventually had wanted to dive into the use of the time indications in the translations, it would have also been convenient to select of highlight the time indications in the fragments, perhaps in a different color, whilst annotating. This sidestep in this research has now been left behind, since it would have cost too much time to go back and annotate all time indications for all 250 fragments in all five languages.

6. Conclusion

In this research, intended as part of the Time in Translation project, I analyzed the ways in which the Present Perfect Continuous (PPC) tense is translated into Dutch, French, German, Spanish and Portuguese. After having extracted, annotated and categorized fragments containing a PPC from Europarl, a multilingual parallel corpus, these fragments were analyzed and their categories visualized with the use of Multidimensional Scaling.

The general research question in this thesis was: which approaches are used most and which work best when it comes to the translation of the Present Perfect Continuous? Most creative solutions in the translations used for this research were in line with the initial expectations of nominalization and usage of time indications, the latter used mostly in combination with the Present tense in all languages. Without the presence of these time indications, however, no creative translation seemed to accurately preserve the continuous meaning of the PPC, except for one surprising Spanish translation which made use of repetition to accurately translate the PPC in question.

The sub question in this thesis was; how do these approaches differ between the French, Dutch, German, Spanish and Portuguese languages? These languages seemed to behave in very similar ways; no big differences were found in the data of this research.

Looking back, some simple configurations to the order of the steps taken in this research could have made the process easier and the results perhaps a little more comprehensive and reliable, but for a small research within a startup project, I would say it has provided us with a more than interesting analysis of the Present Perfect Continuous tense. Further research on the translation on this and other complex tenses with no one-to-one translation will bring us closer to developing machine translators that are able to construct creative translations like a native speaker would.

References

van der Klis, M., Le Bruyn, B., & de Swart, H. (2017). Mapping the Perfect via Translation Mining. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* 2017, 497-502.

Nishiyama, A. and Koenig, J. P. (2010). What is a perfect state? Language, 86, 611-646.

Wälchli, B. & Cysouw, M. (2012). Lexical typology through similarity semantics: Toward a semantic map of motion verbs. In Linguistics 50-3 (2012) (pp. 671-710)

Website of EuroParl corpus: http://www.statmt.org/europarl/

Website of Time in Translation project: http://timealign.pythonanywhere.com/