CAMUS IN KAART

Kieke Eltje Emilie Swager (4099427)
Begeleiders: Henriëtte de Swart, Martijn van der Klis, Bert le Bruyn
Tweede beoordelaar: Johannes Korbmacher
Datum: 21 april 2017
Bachelor Kunstmatige Intelligentie, Universiteit Utrecht
Bacheloreindwerkstuk 7,5 ECTS

1 Inhoudsopgave

2 Inleiding

Aujourd'hui, j'ai terminé ma recherche.
Vandaag, heb ik mijn onderzoek afgerond.
Heute habe ich meine Untersuchung abgeschlossen.
Today, I have completed my research.
Hoy, he completado mi investigación.

Als afsluiting van mijn bachelor en als opmaat naar de praktijk is dit werkstuk uiteindelijk mijn bacheloreindwerkstuk geworden. In het begin was het voor mij best lastig om een keuze te maken voor een onderwerp. De studie Kunstmatige Intelligentie is zeer breed opgebouwd, met een grote spreiding van gebieden waarin vakken worden aangeboden. Het leek me hierdoor verstandig eerst voor mezelf een richting te bepalen waarin ik mijn scriptie wilde gaan schrijven, zodat de keuze voor een onderwerp wat meer beperkt werd.

Ik heb gemerkt dat tijdens mijn studie Kunstmatige Intelligentie mijn interesse vooral is gegroeid in het gebied van taal. Ik heb mijn vrije keuzevakken ingedeeld met vakken uit het taalgebied en dit is mij goed bevallen. Om deze reden besloot ik mij te focussen op onderwerpen binnen dit gebied. Hierbinnen ben ik verder breed geïnteresseerd en heb ik geen sterke voorkeur voor bepaalde onderwerpen. Zowel de syntax als de semantiek interesseren mij zeer. Bovendien heb ik voor mijn gevoel van beide onderwerpen niet heel veel meer dan de basis behandeld, waardoor ik vind dat ik nog niet een duidelijke voorkeur kan uitspreken.

Op Blackboard heb ik de lijst met mogelijke begeleiders en onderwerpen geraadpleegd. Op deze lijst stond prof. dr. Henriëtte de Swart met het onderwerp computertaalkunde. Ik heb een mail gestuurd naar mevrouw de Swart, waarna ik een enthousiaste mail terug kreeg met een korte uitleg over het bestaande onderzoek waar ik deel van uit zou kunnen maken. We hebben een moment afgesproken om bijeen te komen, samen met begeleiders dr. Bert Le Bruyn en Martijn van der Klis.
Tijdens deze bijeenkomst is voor mij globaal meer duidelijk geworden over het bestaande onderzoek en wat de mogelijkheden voor mij zouden zijn om verder onderzoek in te gaan doen. Na de afspraak was ik enthousiast geworden om aan dit taalonderzoek mee te werken en dit onderzoek te koppelen aan mijn bacheloreindwerkstuk.

Ik kon bijdragen aan het hoofdonderzoek door de Perfect in de literatuur te onderzoeken, in het Nederlands, Frans, Duits, Engels en Spaans. De Perfect wordt in deze vijf talen op dezelfde manier gevormd, namelijk met een hulpwerkwoord (‘zijn’ of ‘hebben’) en een voltooid deelwoord. De Perfect in het Nederlands is de Voltooid Tegenwoordige Tijd (VTT) en ziet er bijvoorbeeld als volgt uit: *Jij hebt een boek gelezen*. Het hulpwerkwoord is hier ‘hebt’ en het voltooid deelwoord ‘gelezen’. De VTT kan ook gevormd worden met ‘zijn’ als hulpwerkwoord. Bijvoorbeeld: *Wij zijn geschrokken van het alarm.* ‘Geschrokken’ is het voltooid deelwoord van schrikken. *Tu as lu un livre* is een voorbeeld van een Perfect in het Frans, de Passé Composé. Hier is ‘as’ het hulpwerkwoord, een vervoeging van avoir. ‘Lu’ is het voltooid deelwoord van lire, wat ‘lezen’ betekent.

Mijn begeleiders stelden voor dit onderzoek te doen aan de hand van het boek L’Étranger van de Franse filosoof, journalist en schrijver Albert Camus. Dit voorstel kwam doordat De Swart en Molendijk (2002) al eerder opgevallen is dat L’Étranger een speciale manier van schrijven heeft. Camus maakt in zijn roman namelijk veel gebruik van de Passé Composé, terwijl in Franse romans over het algemeen gebruik wordt gemaakt van de Passé Simple. Hiervan wijkt Camus dus af en door de vele voorkomens van de Passé Composé, is dit boek interessant voor dit onderzoek. Bij eerdere onderzoeken zijn ongeveer 500 voorkomens van de Perfect geanalyseerd. Het is de bedoeling rond dit aantal uit te komen om de

onderzoeken vergelijkbaar te houden. Door het grote aantal voorkomens van de Passé Composé in L'Étranger kan dit aantal snel verkregen worden.

Omdat het onderzoek zich richt op de Perfect in de literatuur, hebben we te maken met gedrukte boeken. De data is het makkelijkst verwerkbaar als de tekst van de boeken digitaal staan. Om dit te bereiken is OCR (Optical Character Recognition) de overwogen optie. Dit is een hulpprogramma om gedrukte tekst digitaal en bewerkbaar te maken. Hoe dit programma werkt, is terug te lezen bij de *Methode.* Binnen het hoofdonderzoek is nooit eerder gebruik gemaakt van deze optie en het is dan ook afwachten of dit makkelijk toepasbaar is. Als deze optie geschikt blijkt, biedt dat een opening naar veel mogelijke vervolgonderzoeken.

Als de teksten digitaal staan, moeten de contexten bepaald worden. De contexten zijn de werkwoorden in de overeenkomende fragmenten die we in iedere taal analyseren. Om onze contexten te verkrijgen, moeten eerst de Perfects uit de Franse tekst geëxtraheerd worden. Dit gebeurt in eerste instantie automatisch, maar de foutief geselecteerde Perfects dienen handmatig aangepast te worden. Vervolgens kunnen de Franse Perfects gekoppeld worden aan de werkwoordsvormen in de vertalingen. Deze moeten handmatig geselecteerd worden. Als deze geselecteerd zijn, kunnen de werkwoordstijden worden toegekend. Aan het Nederlands en Engels worden de werkwoordstijden automatisch toegekend, maar bij het Duits en Spaans moet dit handmatig gebeuren. De Nederlandse en Engelse contexten moeten wel gecontroleerd en, indien nodig, handmatig aangepast worden. Als aan alle contexten een werkwoordstijd is toegekend, kan de data verwerkt worden. Van de data worden semantische kaarten gemaakt. Deze semantische kaarten zijn in het boek *Indefinite Pronouns* van Haspelmath (1997) voor het eerst geïntroduceerd. Dit zijn kaarten waarop punten te zien zijn die de werkwoordstijden van alle contexten visualiseren. Op deze kaarten wordt de verdeling van de werkwoordstijden duidelijk waardoor de resultaten goed te analyseren zijn. Het gehele proces is per stap beschreven onder het kopje *Methode* en hoe dit in de praktijk te werk is gegaan, is beschreven onder het kopje *Uitvoering*.

Mijn onderzoek is een uniek deelonderzoek binnen het hoofdonderzoek. Voor het eerst worden de gedragingen van de Perfect binnen de literatuur onderzocht. Het is interessant om te ondervinden of de Perfect zich anders zal gedragen binnen dit gebied, vergeleken met de gedragingen die reeds zijn gevonden binnen de formele en informele gesproken taal. Het is dus ook voor het eerst dat data wordt verwerkt vanuit gedrukte tekst. Deze data moet nog gedigitaliseerd worden. Het was eerst onzeker of dit zou gaan lukken en het was de vraag of OCR een oplossing zou bieden. In dit verslag is te lezen hoe dat proces is verlopen.
Bovendien is de opzet van mijn onderzoek onderscheidend vergeleken met eerdere onderzoeken. In mijn onderzoek is Frans de enige brontaal waaruit de Perfects zijn geëxtraheerd. Vervolgens heb ik geanalyseerd hoe deze Perfects zich in de andere talen gedragen. Bij eerdere onderzoeken hebben de Perfects verschillende brontalen en zijn de gedragingen van deze Perfects vanuit hun brontaal vergeleken met de overige talen.

De verbinding met de Kunstmatige Intelligentie is in dit onderzoek op twee manieren terug te vinden. Allereerst in de manier waarop de data is verwerkt. Bijvoorbeeld de (bestaande) algoritmen waarmee onder andere de teksten parallel zijn gezet, de Perfect vormen zijn geëxtraheerd en de werkwoordstijden zijn toegekend. Maar de belangrijkste bijdrage van de Kunstmatige Intelligentie wat betreft de dataverwerking, is de methode die is gebruikt om de semantische kaarten te vormen. De semantische kaarten zijn van grote waarde voor dit onderzoek, waardoor de Kunstmatige Intelligentie een onmisbaar onderdeel is.
Eveneens is een verbinding te zien aan de hand van het uiteindelijke resultaat van het onderzoek. De bedoeling van het onderzoek is om meer duidelijkheid te krijgen over de gedragingen van de Perfect. Als hier meer duidelijkheid over komt, kan dit worden teruggekoppeld in bijvoorbeeld machine translation (computervertaling). Met behulp van de resultaten van dit onderzoek kan de narratieve Perfect beter worden vertaald in verschillende talen, waardoor deze vertalingen natuurlijker aan zullen voelen.

3 Achtergrond van het onderzoek

3.1 Eerder onderzoek

Mijn onderzoek is een uitbreiding van het reeds opgezette hoofdonderzoek Time in Translation. Binnen dit onderzoek wordt gekeken naar de gedragingen van de Perfect binnen en tussen verschillende talen. Het is namelijk gebleken dat in verschillende talen de Perfect zich op onderscheidende manieren kan gedragen. De Perfect wordt op verschillende manieren gebruikt, zowel op zinsniveau als in gesprekken. Hierbij kan het moment waar de Perfect zich in bevindt ook verschillen. De Perfect kan gebruikt worden als over het verleden wordt gesproken, maar ook als het over het heden gaat. Hier zijn al eerdere onderzoeken naar gedaan, maar deze waren vaak meer van kwalitatieve aard. In de volgende paragrafen worden de eerdere onderzoeken van Time in Translation beschreven. De bedoeling van dit onderzoek is om, door middel van een combinatie van een kwalitatief en kwantitatief onderzoek, duidelijk te maken wat de betekenis van een Perfect precies inhoudt en hoe deze betekenis tot stand komt. Hopelijk zal tijdens dit deelonderzoek steeds meer duidelijk worden over de betekenis van de Perfect en het ontstaan hiervan.

Het Time in Translation project is begonnen met het onderzoeken van de gedragingen van de Perfect in het formele taalgebruik binnen het Europees Parlement. Vergaderingen van het Europees Parlement worden in alle Europese talen gedigitaliseerd. Deze gedigitaliseerde vergaderingen zijn opgeslagen in het EUROPARL corpus (Tiedemann, 2012), welke getagd en gelemmatiseerd is. Dit betekent dat van alle woorden de woordsoorten zijn aangegeven, en voor ieder woord het lemma is aangegeven. Doordat deze informatie al was toegevoegd aan de data, was het gemakkelijk om bijvoorbeeld de Perfect vormen te extraheren.
De bedoeling van een vertaling is dat de betekenis gelijk blijft. De tijd van de werkwoorden kan wel variëren per taal. Dit is hetgeen wat interessant is voor dit onderzoek.
Bij het EUROPARL onderzoek is naar voren gekomen dat het gebruik van de Perfect zich voornamelijk in het verleden bevindt. Dit lag in de lijn der verwachting, met de gedachte dat tijdens de vergaderingen van het Europees Parlement voornamelijk wordt gesproken over zaken in het verleden. Dit zijn namelijk de situaties in het verleden waar bijvoorbeeld problemen waren, die aangekaart worden zodat deze problemen opgelost kunnen worden.

Als vervolg op dit onderzoek hebben twee medestudenten, Anne en Vincent, onderzoek gedaan binnen het informele taalgebruik. Dit hebben ze gedaan aan de hand van de ondertiteling van films. Van veel films is de ondertiteling in verschillende talen te vinden in het OpenSubtitles corpus (Lison & Tiedemann, 2016), waarin deze al parallel gezet zijn.
Anne en Vincent hebben tijdens hun onderzoek uit vijf films alle Perfects geëxtraheerd, ongeacht uit welke taal de Perfect kwam. De taal waar de Perfect uit geëxtraheerd is, is de brontaal. Vervolgens zijn deze Perfects op dezelfde manier als in het EUROPARL onderzoek gekoppeld met de werkwoorden uit de andere talen. Van deze data zijn semantische kaarten gemaakt, en deze zijn vergeleken met de semantische kaarten van EUROPARL onderzoek.
De verwachting van Anne en Vincent was dat de Perfect in de informele gesproken taal van films zich meer in het heden en de toekomst zou gedragen. Uit de semantische kaarten was af te lezen dat de Perfect zich vaak vertaalde in de tegenwoordige tijd of de toekomende tijd, dus hun hypthese bleek te kloppen.

Mijn onderzoek zal zich richten op de geschreven taal van een vertellend verhaal. We willen graag onderzoeken welke werkwoordstijden in de verschillende vertalingen worden gebruikt en om te proberen het gebruik van die werkwoordstijden te verklaren.

3.2 Literatuur

Mijn hypotheses voor mijn onderzoeksvragen heb ik grotendeels gebaseerd op bestaande literatuur. De literatuur die ik hiervoor heb geraadpleegd, staat hieronder beschreven.

Artikelen

Nishiyama en Koenig (2010) laten zien dat de Engelse Perfect op verschillende manieren geïnterpreteerd kan worden. Bijvoorbeeld in zin (1), met bijbehorende interpretaties (2a) en (2b).

1) Ken has been sick.
2) a. Ken is still sick.
   b. Ken is not sick anymore.

Zin (1) kan zich dus, zoals de interpretaties in (2a) en (2b) laten zien, respectievelijk in het heden of in het verleden bevinden. Op welke manier je zin (1) moet interpreteren, is slechts te bepalen aan de hand van de context of door een zinsdeel die erachter staat en meer duidelijkheid geeft over de tijd van de Perfect. In dit artikel worden de verschillende manieren waarop een Perfect zich kan gedragen uitgelegd.

Reichenbach (1947) heeft een boek geschreven waarin onder andere de Engelse Present Perfect is onderzocht. Hierin heeft hij een Reichenbachiaanse structuur ontwikkeld. Aan de hand van (3a) en (3b) zal ik kort uitleggen hoe deze structuur is opgebouwd.

3) a. Sara left the party. *(Simple Past)*
   b. Sara has left the party. *(Present Perfect)*

In zowel (3a) als (3b) heeft de gebeurtenis (**E**), het vertrekken van Sara, plaatsgevonden voor het moment van spreken (**S**). In (3b) wordt een gebeurtenis in het verleden beschreven, waarbij het moment van spreken van belang is. De zin vertelt ons dat Sara wegging van het feest en op dit moment niet meer op het feest is. Om het verschil tussen de Present Perfect en de Simple Past duidelijk te maken, is naast (**E**) en (**S**) nog een derde notie in het leven geroepen. Dit is de referentietijd (**R**). In (3a) wordt een handeling beschreven, waarbij het niet duidelijk wordt wanneer het resultaat van de handeling plaatsvindt. Hierdoor vallen de gebeurtenis (**E**) en de referentietijd (**R**) samen, aangegeven met (**,**). Beide noties vinden plaats vóór het moment van spreken (**S**), aangegeven met (**-**). Dit wordt weergegeven als **E,R-S**. In (3b) valt de referentietijd (**R**) samen met het moment van spreken (**S**). De gebeurtenis (**E**) vindt plaats vóór deze twee noties, waardoor de structuur voor de Present Perfect **E-R,S** is. Dit wordt de Reichenbachiaanse structuur genoemd.

In het artikel van De Swart (2007) wordt de Perfect onderzocht in het Engels, Frans, Nederlands en Duits. De Swart argumenteert dat de belangrijkste verschillen tussen talen niet liggen op het niveau van de zinssemantiek, maar op het niveau van het grotere geheel en in de kenmerken van het gebruik op dat niveau. Er wordt gesteld dat bovengenoemde vier talen allemaal een Perfect zijn van het type dat Reichenbach heeft beschreven, met de E-R,S structuur. De Swart analyseert de Perfect in de vier talen waarbij de focus ligt op de temporele structuur en de aspectuele eigenschappen. Hierbij is naar voren gekomen dat talen verschillen op het gebied van mogelijke relaties tussen de tijd van de besproken gebeurtenis (**E**) en andere tijden of gebeurtenissen in de zin of de tekst eromheen die met (**E**) te maken hebben.
De Engelse Present Perfect blokkeert alles wat een temporele relatie aangaat met de tijd van de gebeurtenis (**E**). Hierdoor kan de Perfect niet gebruikt worden als een specifieke tijdsbepaling in de zin staat. Bijvoorbeeld in onderstaande zin (4).

4) *Sara has left the party at six o'clock.

Dit is dezelfde zin als (3b), maar nu met een toevoeging van een specifieke tijdsbepaling. Doordat deze tijdsbepaling ervoor zorgt dat de gebeurtenis (**E**) op een bepaald tijdstip wordt geplaatst, is de onwelgevormd geworden. Dit is in overeenstemming met het feit dat specifieke tijdsbepalingen invloed zouden moeten hebben op de referentietijd (**R**), en niet op (**E**). Aangezien (**R**) samenvalt met

(**S**), gaat de Present Perfect alleen samen met aanwijzende temporele bijwoorden, zoals bijvoorbeeld ‘gister’, ‘vanmiddag’ of ‘onlangs’.
Ook is de Engelse Present Perfect geen geschikte werkwoordstijd voor het vertellen van verhalen, omdat het bij verhalen nodig is te kunnen wisselen tussen gebeurtenissen.
De Nederlandse Perfect, de Voltooid Tegenwoordige Tijd (VTT), kan wel worden gebruikt in een zin met een specifieke tijdsbepaling. Zie de welgevormde zin (5).

5) Sara is om drie uur vertrokken.

Daarentegen kan, net zoals in het Engels, de VTT niet gebruikt worden bij het vertellen van verhalen.
De Perfectvormen van het Frans, de Passé Composé, en van het Duits, de Perfekt, zijn makkelijker in gebruik. Deze vormen hebben minder restricties dan de Engelse en Nederlandse. De Perfects in het Frans en Duits kunnen zonder problemen gecombineerd worden met specifieke tijdsaanduidingen en kunnen gebruikt worden op een vertellende manier van taalgebruik.

Het artikel van Nishiyama en Koenig (2010) maakt duidelijk dat het Perfect zich op verschillende manieren kan gedragen. Het Perfect kan zich in het heden of het verleden bevinden.
In het boek van Reichenbach (1947) is duidelijk geworden hoe de Engelse Present Perfect zich onderscheidt van de Simple Past. Hier is een belangrijke structuur uit voortgekomen en het heeft ervoor gezorgd dat de Present Perfect zich heeft neergezet als een typische Perfect.
Een belangrijke conclusie van het artikel van de Swart (2007) is dat de Franse en Duitse Perfect wel te gebruiken zijn bij het vertellen van een verhaal en dat de Engelse en Nederlandse Perfect door bepaalde restricties hier niet geschikt voor zijn.

3.3 Onderzoeksvragen

De bron van mijn data is het boek L’Étranger van Albert Camus. Dit is een boek waarin een verhaal verteld wordt. Hoe zal de Perfect vanuit het Franse verhaal zich gedragen in de vertalingen van het Nederlands, Duits, Engels en Spaans?

De data zal geanalyseerd worden aan de hand van semantische kaarten. Als clustering ontstaat in deze kaarten, dan is het mogelijk hier iets zinnigs over te zeggen. Zal clustering ook ontstaan in de semantische kaarten van mijn onderzoek?

De data die ik zal gaan analyseren komt uit de literatuur. Het gaat hier dus om gedrukte boeken. De benodigde data staat dus nog niet digitaal, terwijl dit wel noodzakelijk is om het te kunnen verwerken. Mijn vraag in relatie tot de dataverwerking is dus: in hoeverre is het mogelijk gedrukte teksten te analyseren?

Eerder onderzoek is gedaan binnen het EUROPARL corpus en het OpenSubtitles corpus. Beide corpora bevatten gesproken tekst, terwijl mijn onderzoek zich richt op geschreven taal. Zal een verschil merkbaar zijn wat betreft de variatie van het gebruik van verschillende werkwoordstijden?

3.4 Hypotheses

Aangezien ik onderzoek doe naar de Perfect vorm in het boek L'Étranger, wat een vertellend boek is, heb ik bepaalde verwachtingen over de gedragingen van de Perfect in de vertalingen. Gebaseerd op de conclusie van de Swart (2007) dat alleen de Franse en Duitse Perfect samen gaan met het vertellend taalgebruik, verwacht ik dat de Passé Composé uit het Frans in het Duits zal worden vertaald met een Perfekt. In het Nederlands zal de Passé Composé zich verdelen tussen de VTT en de OVT. Als er een specifieke tijdsbepaling bij staat, zal het in het Nederlands vertaald worden met een VTT. Op het moment dat er een aanwijzend temporeel bijwoord, zoals 'gister', bij de Passé Composé staat, zal deze in het Nederlands en Engels vertaald worden met respectievelijk een VTT en een Present Perfect. Indien zo'n bijwoord ontbreekt, zal het worden vertaald met respectievelijk een OVT en een Simple Past. Over het Spaans durf ik geen uitspraak te doen, omdat ik daar geen relevante literatuur in het Engels of Nederlands voor heb kunnen vinden.

In lijn met mijn verwachting dat de Franse Passé Composé voornamelijk met een Perfekt vertaald zal worden in het Duits, in het Engels met een Simple Past en in het Nederlands met een Onvoltooid Verleden Tijd, verwacht ik dat in ieder geval deze drie werkwoordstijden een clustering zullen gaan vormen.
Hieromheen is het ook mogelijk dat er nog andere clusters zullen vormen. Zoals beargumenteerd in het artikel van de Swart (2007) gaat de Nederlandse Perfect, de VTT, wel samen met een aanwijzend temporeel bijwoord. Hierdoor verwacht ik dat de Passé Composé, die in de zin staat met zo'n bijwoord, in het Nederlands vertaald zal worden met een VTT. Daardoor verwacht ik dat in de Nederlandse semantische kaart de Voltooid Tegenwoordig Tijd (VTT) ook een cluster zal vormen, omdat in een vertellend verhaal zeker gebruik zal worden gemaakt van aanwijzende temporele bijwoorden, zoals 'gister' of 'vorig jaar'.

De boeken moeten gedigitaliseerd worden voordat ik de data kan gaan analyseren. Tegenwoordig zijn veel mogelijkheden beschikbaar om gedrukte tekst te digitaliseren en bewerkbaar te maken. Een hulpmiddel dat hiervoor zou kunnen werken is OCR. Op internet zijn veel versies beschikbaar van dit hulpmiddel, dus het zal even uitzoeken zijn wat de best werkende versie is. Ik heb zelf nog nooit eerder gebruik gemaakt van OCR, dus ik kan niet met zekerheid zeggen hoe gemakkelijk het zal werken voor dit onderzoek. Mijn verwachting is dat de technologie zich inmiddels op zodanig niveau bevindt, dat het mij zal lukken om een programma te vinden die ervoor zorgt dat de boeken gedigitaliseerd worden en de teksten bewerkbaar. Een aspect wat hierin zal meespelen, is de kwaliteit van de boeken die ingescand moeten worden. Als de inkt van de tekst vervaagd is, of als het lettertype onleesbaar is voor OCR, dan zal dit moeilijkheden opleveren.

Eerdere onderzoeken in de gesproken taal zijn gedaan binnen het het EUROPARL corpus en OpenSubtitles corpus. In beide gevallen gaat het om dialogen. Mijn onderzoek wordt gedaan binnen geschreven tekst, waarbij een verhaal wordt verteld.
Ik verwacht een merkbaar verschil in het aantal verschillende soorten werkwoordstijden. Het lijkt mij aannemelijk dat in een geschreven tekst er bewuster wordt gekozen voor een bepaald taalgebruik, waardoor meer van dezelfde werkwoordstijden gebruikt zullen worden.
Bij dialogen zal er meer variatie zijn van werkwoordstijden, omdat minder bewust wordt nagedacht over de keuze van een bepaalde werkwoordstijd.

3.5 Relatie met Kunstmatige Intelligentie

Gedurende dit onderzoek heb ik gemerkt dat het volgen van de studie Kunstmatige Intelligentie van grote waarde is geweest. De kennis en vaardigheden die ik de afgelopen drie jaar heb opgedaan, heb ik nu namelijk perfect kunnen combineren. Hierdoor was ik in staat dit onderzoek naar behoren uit te voeren.
Tijdens mijn studie heb ik onder andere vakken gevolgd binnen de taalkunde, waardoor ik de nodige kennis van de basisbegrippen heb vergaard en meer weet over de manier waarop onderzoek wordt gedaan binnen dit vakgebied. Tevens heb ik met verscheidene computerprogramma's leren werken en weet ik hoe algoritmes in elkaar steken en wat de mogelijke toepassingen daarvan zijn.
Bij dit onderzoek is het van groot belang de kennis over de taalkunde te kunnen verwerken met de nodige kennis vanuit de computerwereld. Deze combinatie kan behaald worden tijdens de studie Kunstmatige Intelligentie, waardoor studenten van deze studie uiterst geschikte onderzoekers zijn voor dit type onderzoek.

4 Het onderzoek

In dit hoofdstuk wordt eerst onder het kopje *Methode* beschreven wat de aanpak van het onderzoek was en wat hier allemaal voor nodig is geweest. Deze aanpak komt grotendeels overeen met die van het hoofdonderzoek binnen de EUROPARL. Deze stappen zijn ook beschreven in het artikel van Van der Klis, Le Bruyn en De Swart (2017). Alleen het verkrijgen van de teksten is op een andere manier gegaan. Na de *Methode* wordt onder het kopje *Uitvoering* beschreven hoe deze aanpak in praktijk is verlopen. De punten die extra aandacht behoeven, zijn beschreven onder het kopje *Aandachtspunten uitvoering*. De uitvoering van het onderzoek heeft gezorgd voor resultaten, die onder het kopje *Resultaten* op een rijtje zijn gezet in de vorm van semantische kaarten en tabellen. Eveneens zijn de resultaten descriptief beschreven. Aan de hand van de resultaten zijn als antwoord op de onderzoekvragen conclusies getrokken. Deze conclusies zijn terug te lezen onder het kopje *Conclusie*. Bij de *Discussie* wordt besproken welke zaken beter onderzocht kunnen worden en worden de conclusies nog eens kritisch besproken.

4.1 Methode

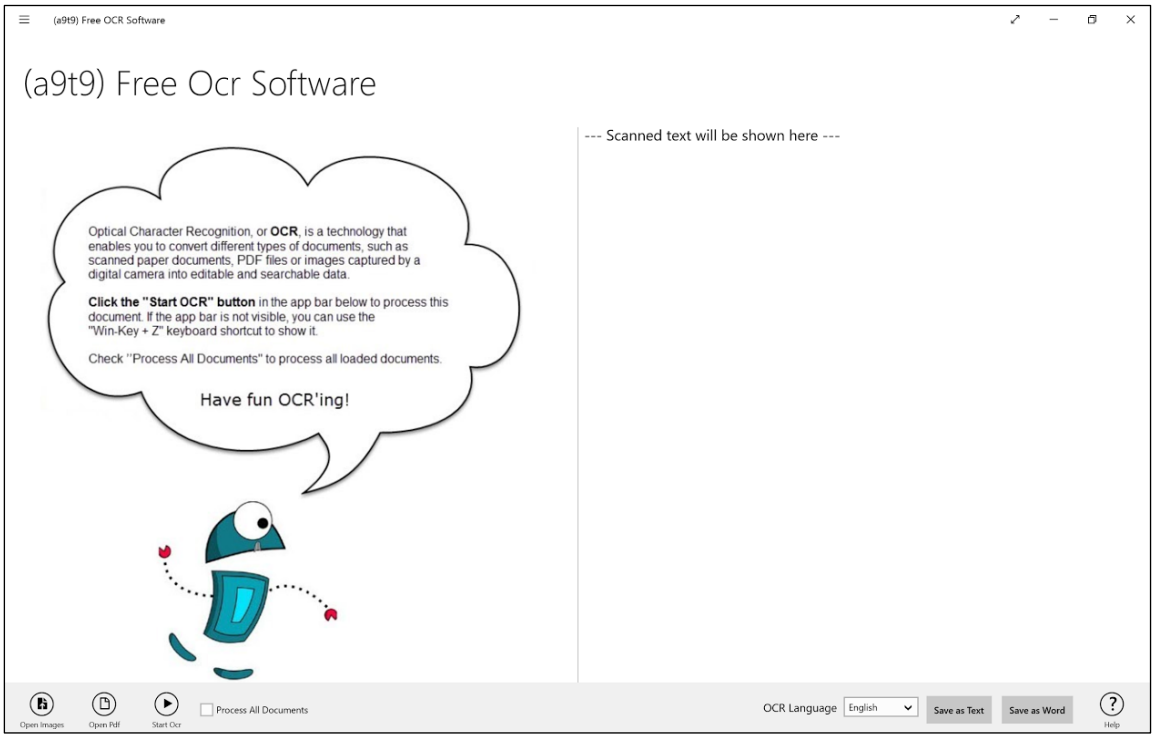
Onder dit kopje wordt de aanpak van het onderzoek uiteengesteld. Per onderdeel wordt uitgelegd wat het precies inhoudt en hoe het gebruikt dient te worden.

De boeken
Voor dit onderzoek is gekozen voor een boek van de Franse schrijver Albert Camus. De Swart en Molendijk (2002) hebben eerder al onderzoek gedaan naar de voorkomens van de Perfects in het boek L'Étranger van Camus. Toen is gebleken dat een groot aantal Perfects wordt gebruikt in het orgineel, vandaar de keuze mijn onderzoek aan de hand van dit boek te doen, samen met de vertalingen in het Nederlands, Duits, Engels en Spaans. De titels van de gebruikte vertalingen zijn respectievelijk 'De Vreemdeling', 'Der Fremde', 'The Outsider' en 'El Extranjero'.

Scanner
Voor het digitaliseren van de boeken is het kopiëren van de boeken de eerste stap. Hierbij is het nodig om een scanner tot je beschikking te hebben die ook een kopieeroptie heeft. Voor het kopiëren is het handig als het contrast en de scherpte in te stellen zijn. Omdat boeken meestal niet in A4 formaat worden uitgebracht en dit wel het formaat is waar de kopieën op worden uitgeprint, is het handig als het apparaat ook kan uitvergroten. Door middel van uitvergroten kan je ervoor zorgen dat de tekst van het boek wordt uitgeprint op het gehele oppervlakte van het A4-papier. Het is de bedoeling dat de tekst zo groot mogelijk is, zodat OCR de tekst beter kan verwerken. OCR is een hulpmiddel dat onder het volgende kopje wordt beschreven.
Nadat de gewenste tekst is gekopieerd, moet deze ingescand worden. Dit doe je door de gekopieerde tekst in de juiste volgorde in de scanner te leggen en het vervolgens te scannen. De instellingen en te volgen stappen hiervoor verschillen per scanapparaat. Vaak kan een e-mailadres worden ingevoerd, waar de gescande documenten als een PDF-bestand naar toe worden gestuurd. Het is handig om een PDF-bestand te hebben van het gekopieerde boek, omdat in de volgende stap OCR programma's vaak de voorkeur geven aan het verwerken van een PDF-bestand.

OCR
Optical Character Recognition is een hulpmiddel waarbij geschreven tekst wordt omgezet in tekens. Deze tekens worden met een programma opgeslagen, waardoor ze bewerkbaar zijn op de computer. Op internet zijn heel veel verschillende programma's voor dit hulpmiddel beschikbaar, zowel betaalde als gratis versies. Deze programma's werken vaak via de website, maar er zijn ook programma's die je apart moet downloaden. Om het programma te laten werken, moet je een PDF-

bestand uploaden, waarna de gedrukte tekst hierin verwerkt zal worden. Tijdens de verwerking wordt de gedrukte tekst omgezet in tekens die je later kunt bewerken, bijvoorbeeld in een tekstverwerkingsprogramma zoals Word.
De Nederlandse, Franse, Duitse en Engelse teksten zijn verwerkt op een Windows computer met het programma a9t9 Free OCR Software. Rechtsonder kan je bij 'OCR Language' de taal kiezen waarin het gewenste bestand verwerkt moet worden. Vervolgens kan je een PDF-bestand importeren door linksonder op de knop 'Open PDF' te klikken. Je kiest het gewenste bestand en klikt op 'Start OCR', waarna deze wordt verwerkt door het programma. Als de tekst verwerkt is, kan je het opslaan als een tekstbestand of als een Word-bestand, door rechtsonder de gewenste manier te kiezen.



Afbeelding 1: Beginscherm a9t9 Free OCR Software

De Spaanse tekst is verwerkt op een MacBook met behulp van het programma van www.onlineocr.net. Een nadeel van deze website was dat het grote bestanden niet in één keer kon verwerken, waardoor dit in gedeeltes moest. Het kostte hierdoor meer moeite om de verwerkte teksten weer bij elkaar te krijgen, maar het werkte wel en daar gaat het uiteindeijk om. Voor een MacBook heb ik niet een beter werkend programma kunnen vinden. A9t9 is niet beschikbaar voor een MacBook, vandaar de gedwongen keuze om een ander programma te gebruiken.
Als alle teksten gedigitaliseerd zijn en deze per taal in een Word-bestand zijn gezet, kunnen deze in de volgende stap parallel gezet worden.

Teksten parallel zetten

Voordat de teksten uit de vijf Word-bestanden verwerkt kunnen worden, moeten deze allemaal parallel gezet worden. Het parallel zetten is een proces dat uit drie stappen bestaat.
Allereerst worden de teksten getokeniseerd op het niveau van hoofdstukken, alinea's, zinnen en woorden. Hiervoor is van de software Uplug (Tiedemann, 2003) de module pre/<taalafkorting>/basic gebruikt. Bijvoorbeeld pre/nl/basic was de gebruikte module voor het Nederlands. Deze module kan platte tekst omzetten naar een XML-formaat, waarin hoofdstukken, alinea's, zinnen en woorden de elementen zijn. Voor iedere taal worden regels gebruikt om het einde van zinnen en woorden te bepalen. Een punt is bijna altijd een zins- en woordeinde, maar niet als deze punt wordt gebruikt als

deel van een afkorting. Dit is een voorbeeld van een uitzondering waar rekening mee gehouden wordt door de module.
Nadat de teksten getokeniseerd zijn, kunnen de documenten gealigneerd worden op zinsniveau met de module align/hun. De aligneermodule maakt gebruik van de software hunalign (Varga et al., 2005). Dit aligneren zorgt voor bestanden waarin per taalpaar wordt aangegeven welke zinnen vertalingen van elkaar vormen.
Als laatste stap wordt er met TreeTagger (Schmid, 2013) part-of-speech tags, dat zijn woordkenmerken zoals 'zelfstandig naamwoord', en lemmata, dat zijn de vormen van de woorden zoals ze in het woordenboek staan, aan de getokeniseerde tekst toegevoegd. Deze informatie wordt gebruikt bij het extraheren van de Perfects.
Vervolgens wordt gebruik gemaakt van een klein Python-script (treetagger-xml, 2017) om de resultaten van TreeTagger te koppelen aan de eerder getokeniseerde bestanden van Uplug. Het koppelen van de resultaten van de TreeTagger aan de getokeniseerde bestanden van Uplug zou Uplug zelf ook moeten kunnen (module pre/<taalafkorting>/all-treetagger), maar op het moment van schrijven blijkt deze niet naar behoren te functioneren. Hierdoor is gekozen om gebruik te maken van een Python-script die deze stap kan uitvoeren.
Op het moment dat al deze stappen zijn uitgevoerd, is er per taal een bestand met getokeniseerde tekst met part-of-speech tags en lemmata, en per taalpaar een bestand met daarin aangegeven welke zinnen vertalingen van elkaar vormen.
Deze bestanden kunnen gebruikt worden in de volgende stap, het extraheren van Perfects met behulp van de PerfectExtractor.

PerfectExtractor

Met deze toepassing kunnen de Perfects uit een gewenste tekst worden herkend. Hierdoor kunnen de fragmenten die een Perfect bevatten los van elkaar bekeken worden, waardoor deze makkelijker onderzocht en vergeleken kunnen worden met de werkwoordsvormen van overeenkomende contexten in de andere talen.
Hoe de PerfectExtractor precies werkt, is terug te lezen in het bacheloreindwerkstuk van Verkleij en Wimmers (2016). Voor het Nederlands, Duits, Frans, Engels en Spaans zijn algoritmes beschikbaar om deze Perfects te extraheren. Deze algoritmes zijn eerder ontwikkeld voor het onderzoek naar Perfects in EUROPARL. De code van dit algoritme is te verkrijgen via de website van TimeAlign.

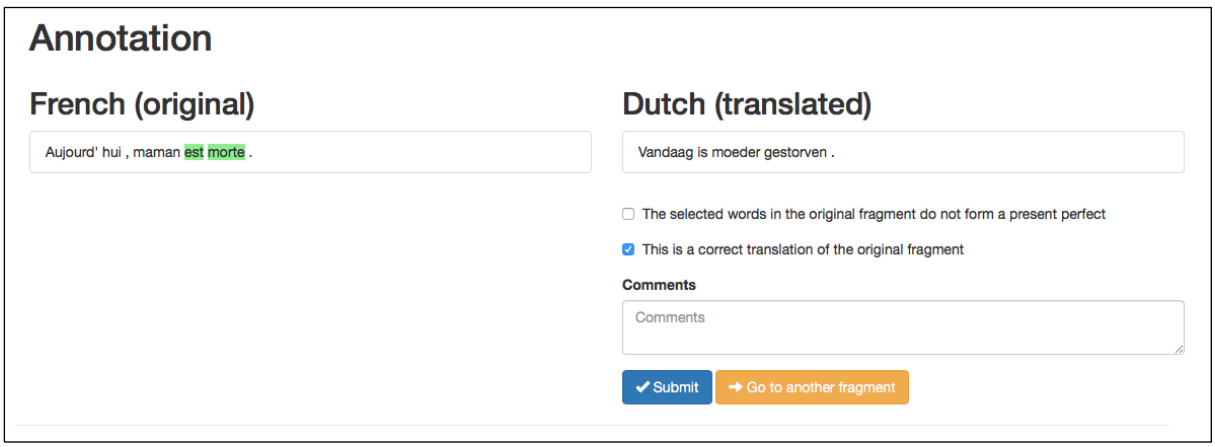
VPSelect

VPSelect is een programma waarmee handmatig werkwoordsvormen geselecteerd kunnen worden. In mijn onderzoek is het selecteren van de Perfects in het Frans automatisch gegaan, doordat de PerfectExtractor de Perfects al uit de Franse tekst had gehaald. Als voor een bepaalde taal geen algoritme bestaat om de Perfects eruit te halen, dan kan VPSelect gebruikt worden om de Perfects te selecteren.
Op het moment dat een context foutief was gemarkeerd als Perfect, dan is dit handmatig aangepast met VPSelect. Tijdens het annoteren kwam ik erachter dat in het Frans vijf contexten foutief als Perfect waren gemarkeerd. Deze contexten heb ik tijdens het annoteren aangegeven door een vakje aan te vinken, die aangeeft dat hetgeen gemarkeerd is in de brontaal geen Perfect is. Dit zal duidelijker worden beschreven bij de *Uitvoering*. Doordat voor deze optie een filter beschikbaar is, kan je alle contexten die zijn opgeslagen met deze optie later gemakkelijk terug vinden.
De contexten die foutief gemarkeerd waren zijn handmatig verbeterd. Doordat mijn onderzoek uitgaat vanuit één brontaal, het Frans, waren alle foutieve Perfects verbeterd en ben ik later bij het annoteren van de andere talen geen foutieve Perfect meer tegengekomen.

TimeAlign

TimeAlign is het programma dat overeenkomende fragmenten uit de brontaal en vertalingen naast elkaar laat zien, zodat de gemarkeerde contexten uit de brontaal (original) gekoppeld kunnen

worden aan de werkwoordsvormen in de vertalingen (translated). In voorgaande stappen is ervoor gezorgd dat alle Perfects in het Frans gemarkeerd waren. Hierdoor kon ik met TimeAlign in de vertalingen de werkwoordsvormen selecteren die daarmee overeenkwamen. Zie afbeelding 2 voor een voorbeeld. Hoe ik gebruik heb gemaakt van TimeAlign, is uitgebreid beschreven onder het kopje *Uitvoering*, bij *Annoteren*.



Afbeelding 2: Voorbeeld TimeAlign fragment (FR-NL)

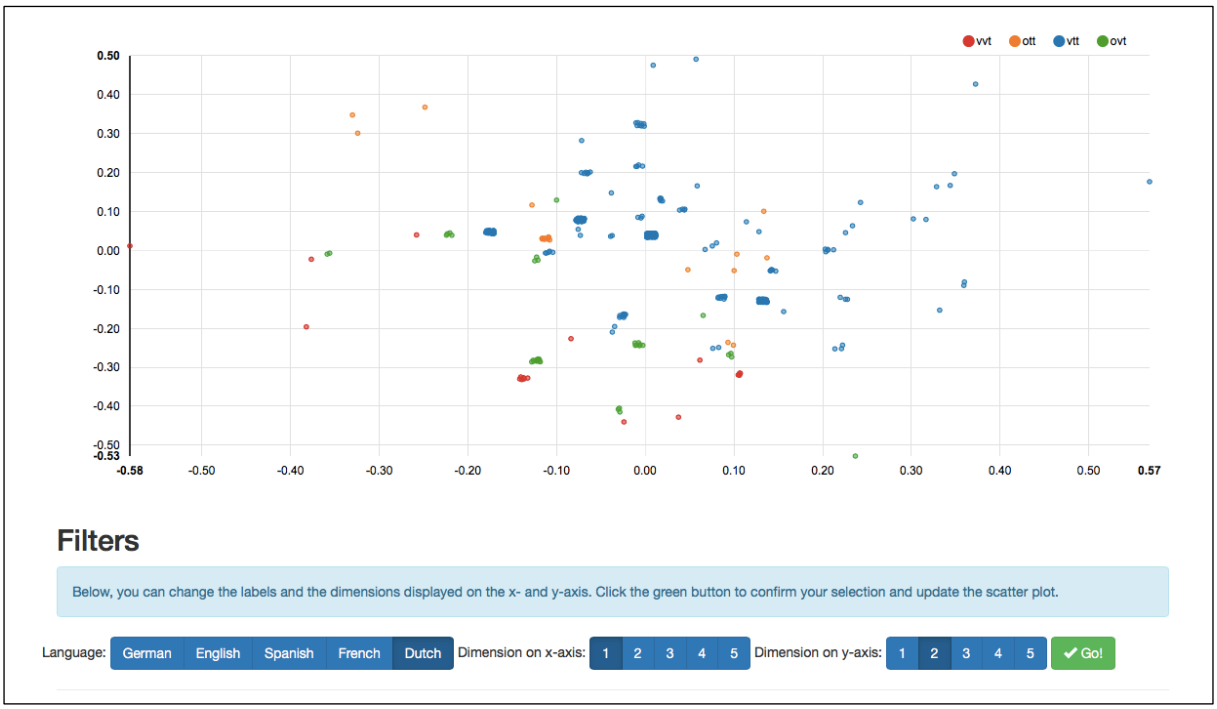
Werkwoordstijden toevoegen

Als de werkwoordsvormen gelijk zijn gesteld aan de contexten van de brontaal, is het nodig de tijd van de geselecteerde werkwoordsvormen te taggen. Voor het Frans was dit niet nodig, want dit was in alle gevallen de Passé Composé. Voor het Engels en het Nederlands werd dit automatisch gedaan met behulp van een algoritme, die ook eerder al ontwikkeld is voor het EUROPARL onderzoek. Doordat er voor het Duits en het Spaans nog geen algoritmes beschikbaar zijn, moesten deze tijden handmatig toegekend worden. Er zijn hier nog geen goede algoritmes voor geschreven omdat in deze talen ambigue vormen voorkomen. In het Spaans kan tomamos de wij-vorm zijn van de presente (tegenwoordige tijd) – *wij drinken* -, maar ook de wij-vorm van de pretérito perfecto simple (verleden tijd) – *wij dronken*. Dit is (nog) niet te onderscheiden met behulp van een algoritme, maar vaak is het wel mogelijk te bepalen welke vorm het is aan de hand van context of door woorden zoals 'ayer', wat 'gister' in het Spaans betekent. Dan is het duidelijk dat je met een verleden tijd te maken hebt. In de *Uitvoering* is beschreven hoe het handmatig toevoegen van de werkwoordstijden is gedaan.

TimeMapping

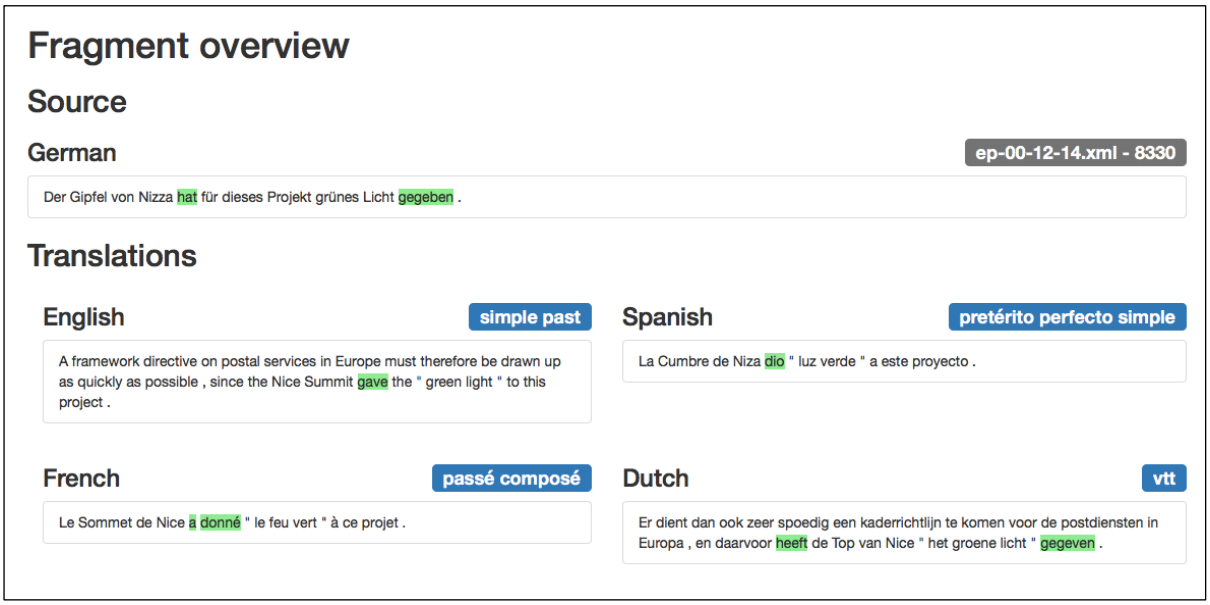
De Passé Composé in het Frans wordt gekoppeld aan de gemarkeerde werkwoordsvormen van het Nederlands, Duits, Spaans en Engels. Hierdoor ontstaat een 5-tupel, waarin wordt aangegeven welke tijd de gemarkeerde context heeft. Het 5-tupel heeft een vaste volgorde van de talen, namelijk *<Nederlands, Frans, Duits, Engels, Spaans>*. Een 5-tupel kan er als volgt uitzien: *<Voltooid_Tegenwoordige_Tijd, Passé_Composé, Präteritum, Simple_Past, Pretérito_Perfecto_Compuesto>.* Omdat in mijn onderzoek enkel vanuit de Perfect in het Frans gekeken wordt, zal bij alle 5-tuples altijd *Passé_Composé* op de tweede plek staan. De tijden van de andere talen zullen wel variëren en in deze variatie zijn we geïnteresseerd.
De 5-tupels worden verwerkt met behulp van MDS (Multidimensional scaling), op dezelfde manier als dat Wälchli en Cysouw (2012) dat hebben gedaan in hun onderzoek. Van de 5-tupels is het mogelijk een matrix te vormen met daarin de afstanden tussen de tupels. Deze afstand is gedefinieerd door twee tupels te vergelijken. Als de leden van de beide tupels precies hetzelfde zijn, hebben ze afstand 0. Als één werkwoordstijd binnen de twee tupels afwijkt, dan geeft dit een afstand 1, gedeeld door het aantal leden van de tupel, vijf in het geval van mijn onderzoek. Hierdoor is de afstand dan 1/5. De afstanden in de matrix worden vervolgens verwerkt met behulp van MDS.

Hiervoor is gebruik gemaakt van het scikit-learn package (Pedregose et al., 2011). Dit is een Python package voor machine learning. De resultaten zijn gevisualiseerd met het nvd3 package (2017). Doordat de werkwoordstijden hun eigen label hebben, is duidelijk te zien hoe een werkwoordstijd zich gedraagt. De overeenkomende werkwoordstijden hebben in elke taal dezelfde kleur, waardoor de kaarten ook makkelijk naast elkaar te vergelijken zijn.
Op de website van TimeAlign zijn de semantische kaarten van het EUROPARL onderzoek terug te vinden. In afbeelding 3 is een voorbeeld te zien van een semantische kaart. Hierin zijn alle datapunten van het Nederlands in dimensie 1 en 2 te zien. Als je je muis op een datapunt laat staan, krijg je te zien welke 5-tupel op dat punt geplaatst is. Onderin kan je met de blauwe knoppen kiezen van welke taal je een kaart wilt zien. Ook zijn hier de dimensies van de x-as en y-as in te stellen. Als de gewenste instellingen zijn gekozen en je klikt vervolgens op de groene Go! knop, dan wordt met de gekozen instellingen een nieuwe semantische kaart gevormd. Bij de lagere dimensies worden zo veel mogelijk punten al op de juiste plek geplaatst. Als je hoger in de dimensies gaat, dan wordt er minder verklaard over de data. De data wordt zoveel mogelijk 'uit elkaar getrokken' bij dimensies 1 en 2, dit gebeurt minder bij dimensies 3 en 4.



Afbeelding 3: Voorbeeld semantische kaart EUROPARL

Als je met je muis op een datapunt klikt, word je doorgestuurd naar een volgend scherm. Het scherm dat je dan te zien krijgt, is een voorbeeld van te zien in afbeelding 4. Hier staat bovenaan de context in de brontaal, met daaronder de vier vertalingen. Bij de brontaal is in het grijs aangegeven uit welk bestand het frament afkomstig is. Bij de vertalingen is in het blauw aangegeven welke werkwoordstijd is toegekend aan de context.



Afbeelding 4: Voorbeeld informatie van datapunt EUROPARL

In het kort

Voor het onderzoek heb je allereerst een tekst nodig waarvan je het gebruik van de Perfects wilt analyseren. Als deze tekst gedrukt is, is het nodig deze te digitaliseren met behulp van OCR. Als alle teksten gedigitaliseerd zijn en parallel zijn gezet, kun je vervolgens hier de Perfects automatisch uithalen met de PerfectExtractor. De Perfects die niet herkend konden worden door de PerfectExtractor kunnen handmatig geselecteerd worden met VPSelect. De geëxtraheerde Perfects moeten vervolgens gekoppeld worden aan de overeenkomende werkwoorden in de vertalingen. Het selecteren van deze werkwoorden gebeurt met TimeAlign. Voor het Nederlands en het Engels worden deze geselecteerde werkwoorden automatisch een werkwoordstijd toegekend, voor het Spaans en het Duits moet dit handmatig gedaan worden. Als aan alle werkwoorden de werkwoordstijden zijn toegevoegd, kunnen de semantische kaarten gemaakt worden. Dit is de laatste stap en wordt gedaan met TimeMapping.

4.2 Uitvoering

Onder het vorige kopje heb ik op een rijtje gezet welke (hulp)programma's allemaal nodig zijn voor het uitvoeren van het onderzoek en hoe deze programma's werken. Onder dit kopje zal ik vertellen hoe de uitvoering van het onderzoek bij mij in de praktijk is verlopen.

De boeken

Mijn begeleiders hadden het Franse orgineel en de Engelse, Nederlandse en Duitse vertalingen al klaarstaan in de boekenkast. De Spaanse vertaling heb ik besteld via een Spaanse website voor boeken.

Tekst digitaliseren

Net zoals in voorgaande onderzoeken was het de bedoeling om ongeveer 500 contexten te kunnen vergelijken in de verschillende vertalingen. Contexten zijn hetgeen wat we willen onderzoeken, de Perfects in dit geval. In overleg hebben mijn begeleiders en ik eerst besloten dat we de eerste twee hoofdstukken van L'étranger van Albert Camus zouden gaan onderzoeken, met de bijbehorende vertalingen in het Nederlands, Duits, Engels en Spaans. Terwijl ik bezig was met het inscannen van de

Nederlandse en Engelse vertaling, hebben we besloten dat het beter zou zijn als we drie hoofdstukken zouden onderzoeken, om zeker te zijn dat we genoeg contexten zouden hebben.
In het begin was het even uitproberen wat de beste manier was om de teksten vanuit de boekjes digitaal te krijgen. Eerst ben ik begonnen met het handmatig overtypen van de Spaanse vertaling. Dit bleek erg veel tijd te kosten, dus de eerder overwogen optie om de boekjes in te scannen met OCR (Optical Character Recognition) bleek het proberen waard.

Voordat gebruik kan worden gemaakt van OCR moet de gedrukte tekst gekopieerd en ingescand worden. Het begint met het zoeken naar de juiste instellingen voor het kopieerapparaat om de teksten uit de boekjes te kopiëren. Het is de bedoeling om de tekst zo groot mogelijk op een A4-formaat te krijgen, met het contrast zo hoog mogelijk en de tekst zo scherp mogelijk. Als de juiste instellingen zijn gevonden, kunnen alle pagina's tot en met hoofdstuk drie gekopieerd worden. Nadat alles gekopieerd is, kunnen de stapeltjes gekopieerde vertalingen per taal ingescand worden en deze per taal in een PDF-bestand opslaan. Deze PDF-bestanden heb ik vervolgens door OCR laten omzetten in bewerkbare tekst. Hierbij is de kans heel klein dat de bewerkbare tekst foutloos wordt omgezet door OCR. In de gescande bestanden is het namelijk niet te voorkomen dat er printstrepen in de tekst zitten waardoor OCR de zinnen waar deze strepen in de buurt staan niet kan omzetten. Ook worden vlekjes in de tekst soms gedetecteerd als een komma of apostrof. Hierdoor is het nodig om de tekst die omgezet is door OCR te controleren en handmatig te verbeteren, daar waar de tekst niet overeenkomt met de tekst uit het boek.
De Nederlandse, Franse, Duitse en Engelse vertalingen zijn omgezet met het programma a9t9. De Spaanse vertaling is omgezet via een website. Het probleem met deze website was dat je niet te grote bestanden in één keer kon uploaden, waardoor ik het PDF-bestand eerst moest splitten via een PDF split programma (2017) en vervolgens deze pagina's per stuk heb omgezet met OCR. De losse teksten heb ik samengevoegd in één Word bestand, waarna ik de tekst handmatig kon verbeteren. Voor het parallel zetten van de teksten was het belangrijk om de alinea's gelijk te houden. De alinea's zijn in de boeken van de vertalingen over het algemeen hetzelfde gelaten, dus het is handig om de alinea's in te voegen in de Word-bestanden aan de hand van de boeken.

Teksten parallel zetten

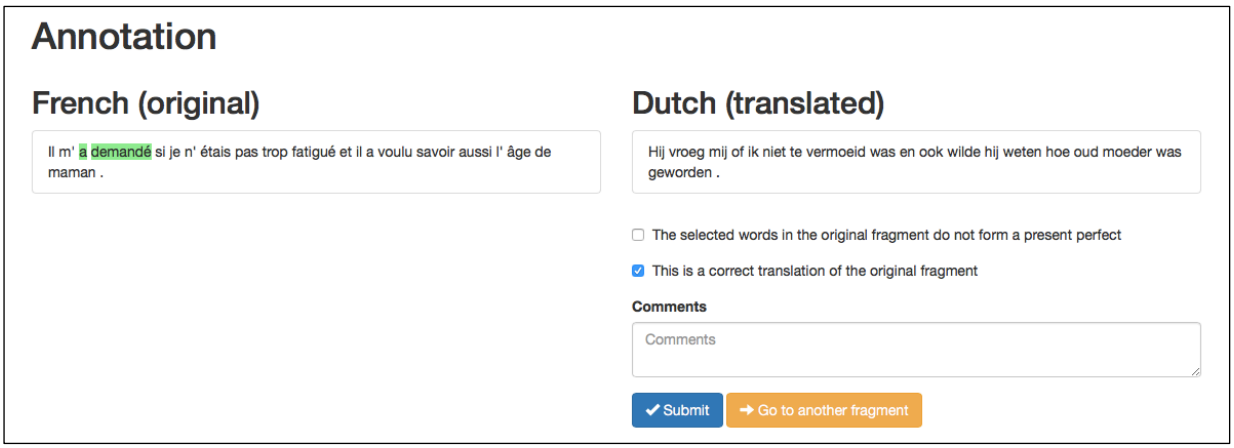
De Franse tekst en de vier vertalingen zijn parallel gezet volgens de stappen beschreven in de *Methode*.

Perfects extraheren

Uit de Franse tekst zijn alle Perfects geëxtraheerd met behulp van de PerfectExtractor. Hierbij wordt de opbouw van de ingevoerde tekst geanalyseerd door de PerfectExtractor. Er wordt gezocht naar kenmerken van een Perfect. Het herkennen van deze kenmerken is geïntergreerd in het algoritme van de PerfectExtractor. Als de kenmerken worden herkend, dan wordt dit gemarkeerd als een Perfect. Met alle geëxtraheerde Perfect vormen kan er verder gewerkt worden.

Annoteren

Het annoteren gaat met behulp van TimeAlign. Dit is een overzichtelijk opgesteld programma, waarmee je in een aantal stappen alle Perfect contexten van de brontaal gelijk kunt stellen aan de overeenkomende werkwoordsvormen in een andere taal. Tijdens het annoteren krijg je telkens een willekeurige Franse context te zien, met daarnaast de zin uit de vertaling die parallel is gezet met de Franse tekst.
In afbeelding 5 is een voorbeeld te zien van een gemarkeerde Franse context, met daarnaast de Nederlandse tekst waarin de overeenkomende context nog geselecteerd moet worden.



Afbeelding 5: ongeannoteerde context (FR-NL)

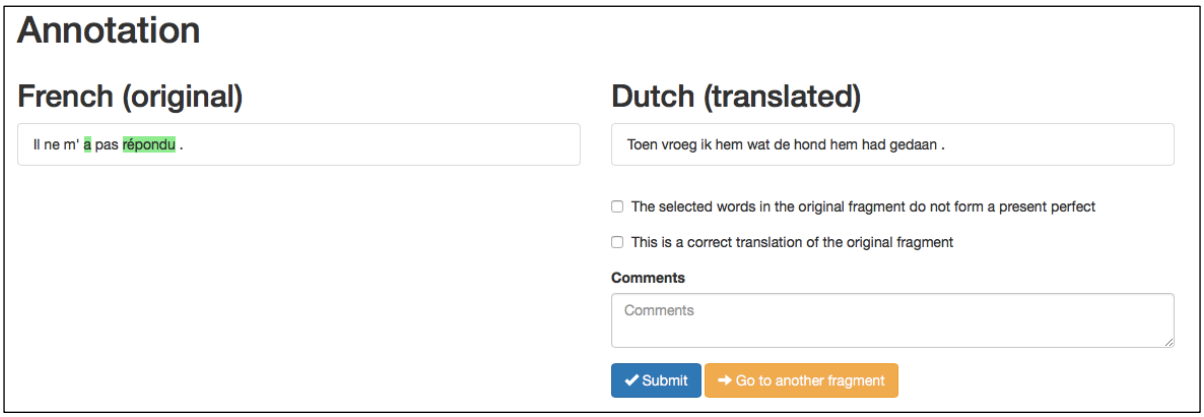
Na het analyseren van de vertaling, moet besloten worden welk werkwoord, of welke werkwoorden, overeenkomen met de gemarkeerde context in de brontaal. Als dit duidelijk is, kan dit aangegeven worden door in de vertaling op het werkwoord, of meerdere werkwoorden, te klikken met de muis. De volgorde van het selecteren maakt hier geen verschil. Door op een woord te klikken zal het geselecteerde woord groen gemarkeerd worden, net zoals de gemarkeerde woorden in de brontaal. Zie afbeelding 6 voor een voorbeeld van een volledig geannoteerde context in de Nederlandse vertaling.

Afbeelding 6: Geannoteerde context (FR-NL)

Nadat de correcte context is geannoteerd in de vertaling, dien je deze te bevestigen door op de blauwe 'Submit' knop te klikken. Door op deze knop te klikken, wordt je antwoord verwerkt en opgeslagen in de database. Hierna zet TimeAlign automatisch de volgende willekeurige context klaar om te annoteren.
Het kan voorkomen dat de PerfectExtractor een foutieve Perfect heeft gemarkeerd. Bijvoorbeeld als een vorm van het hulpwerkwoord 'avoir' is geselecteerd, maar hierbij geen of een foutief voltooid deelwoord is geselecteerd. Het geheel is dan een combinatie van woorden die geen Passé Composé vormen. In dit geval moet je het bovenste vakje aanvinken. 'The selected words in the original fragment do not form a present perfect'. Hiermee geef je aan dat de gemarkeerde context in de brontaal niet een correcte Perfect weergeeft. Het is niet nodig om in de vertaling de woorden te selecteren die de vertaling zouden zijn van de eventueel correcte Perfect, maar dit mag wel. Als je vervolgens op de blauwe 'Submit' knop klikt, zal dit verwerkt worden en opgeslagen in de database. De contexten die op deze manier zijn opgeslagen, zijn later gemakkelijk terug te vinden met behulp van een filter. Met een filter kunnen alle contexten die zijn opgeslagen met bovengenoemde optie

worden weergegeven in een lijst. De fragmenten waarbij wel een Perfect in de zin staat, maar die via de automatische handeling niet correct geselecteerd waren, kunnen vervolgens handmatig worden aangepast met VPSelect. Dit markeren gebeurt op dezelfde manier, door op de correcte woorden te klikken. Op dit moment kun je ook de overeenkomende woorden in de vertaling selecteren, als dat nog niet was gedaan.
Nu de correcte Perfect is geselecteerd met bijbehorende vertaling, is het belangrijk om de bovenste optie uit te vinken, zodat deze context alsnog wordt meegenomen in de analyse. Alle contexten waarbij de bovenste optie is aangevinkt, worden namelijk niet meegenomen in de resultaten.
Het kan ook voorkomen dat in het fragment helemaal geen Perfect voorkomt. Dan moet je niks selecteren in de vertaling en de bovenste optie aanvinken. Vervolgens klik je op de blauwe 'Submit' knop. Hierdoor zal dit fragment niet worden meegenomen in de resultaten.

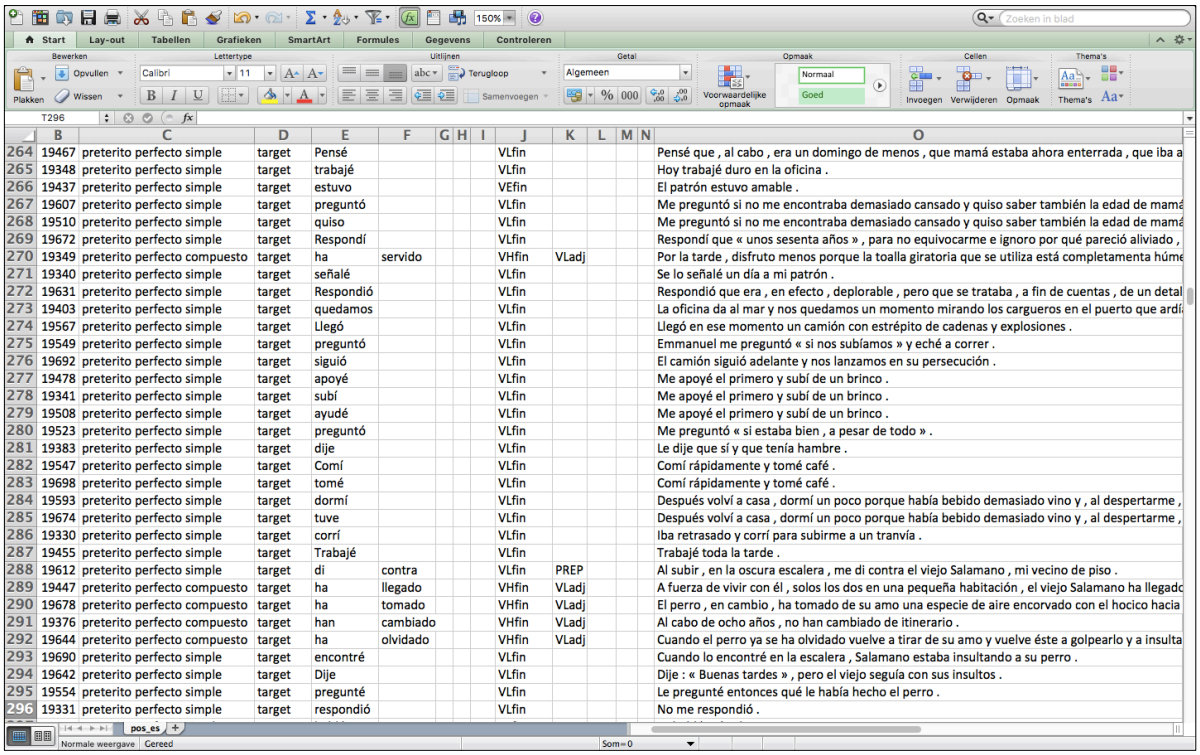
Ook kan het voorkomen dat de context niet op de juiste manier vertaald wordt. Bijvoorbeeld als deze heel vrij wordt vertaald, waarbij de letterlijke betekenis wegvalt. Voor deze gevallen is er ook een optie om dit aan te geven. Dit is de onderste optie die stelt: 'This is a correct translation of the original fragment'.
Deze optie staat standaard aangevinkt, omdat over het algemeen wél een correcte vertaling van de Perfect in de vertaling staat. Als dit niet het geval is, is er de mogelijkheid om deze optie uit te vinken. Als je te maken hebt met een vrije vertaling, dien je het onderste vakje uit te vinken. In dit geval kan je wel de woorden in de vertaling selecteren die de vrije vertaling aangeven. Vervolgens klik je op de blauwe 'Submit' knop. Hierdoor zal de context verwerkt en opgeslagen worden en zal deze gemakkelijk terug te vinden zijn met behulp van het filter. In het geval van een vrije vertaling kan later beoordeeld worden in hoeverre de geselecteerde woorden te vrij vertaald zijn. Als de betekenis voldoende overeenkomt met de betekenis van de brontaal, kan besloten worden om de context alsnog mee te nemen in de analyse. Als dit besloten wordt, dien je de optie weer aan te vinken. Anders wordt de context niet meegenomen bij het verwerken van de data.
Het kan ook voorkomen dat in de vertaling helemaal geen vertaling staat van de Perfect uit de brontaal, dus ook geen vrije vertaling ervan. Dit kan komen doordat de vertaler een fout heeft gemaakt, of doordat het hunalign algoritme een fout heeft gemaakt bij het aligneren.
In afbeelding 7 staat een voorbeeld van een fragment waarbij het niet mogelijk is om in de vertaling woorden te selecteren die enigzins overeenkomen met de context uit de brontaal. In dit geval dien je wederom het onderste vakje uit te vinken. Aangezien er geen juiste vertaling is, kan je vanzelfsprekend geen woorden selecteren in de vertaling, dus deze zal zonder markering blijven. Vervolgens klik je op de blauwe 'Submit' knop.
In het voorbeeld van afbeelding 7 staat de volgende Franse zin: 'Il ne m'a pas répondu'. Dit zou vertaald kunnen worden met zoiets als: 'Hij heeft mij niet geantwoord'. In de Nederlandse vertaling van het fragment in afbeelding 7 staat: 'Toen vroeg ik hem wat de hond hem had gedaan'. Deze vertalingen komen totaal niet overeen. Om deze reden dien je geen woorden te selecteren en de onderste optie uit te vinken, zoals te zien in het voorbeeld.



Afbeelding 7: Juiste vertaling ontbreekt (FR-NL)

Werkwoordstijden handmatig toevoegen

In het Frans was de werkwoordstijd van alle contexten de Passé Composé. Aan de Engelse en de Nederlandse contexten werden de werkwoordstijden automatisch toegevoegd. Voor het Duits en het Spaans was het nodig de werkwoordstijden handmatig toe te voegen. Voor het Duits kon een werkwoordstijd gekozen worden uit de volgende zes werkwoordsvormen: Präsens, Perfekt, Präteritum, Plusquamperfekt, Futur I en Futur II. Voor het Spaans was dit rijtje nog wat langer, er kon namelijk gekozen worden uit vijftien verschillende vormen: presente, pasado reciente, participio, pretérito perfecto compuesto (Perfect), pretérito perfecto simple (Simple Past), pretérito pluscuamperfecto, pretérito imperfecto, condicional simple, condicional compuesto, futuro simple, futuro compuesto, futuro imperfecto, futuro perfecto, infinitivo simple, infinitivo compuesto. Vooral bij het Spaans was het belangrijk om het toevoegen van de tijden aandachtig te doen. In het Spaans heb je namelijk werkwoordsvormen die ambigu zijn. Bijvoorbeeld de werkwoordsvorm 'tomamos', zoals bovenstaand ook genoemd bij het kopje *Methode.* Hierbij is het belangrijk dat de zin waar het werkwoord in staat of de context goed geanalyseerd worden en aan de hand daarvan te bepalen welke tijd toegekend moet worden aan het werkwoord.
In afbeelding 8 is een deel van het Excel-bestand te zien waarin de werkwoordstijden handmatig toegevoegd moesten worden. In kolom C zijn de werkwoordstijden nu handmatig toegevoegd, maar deze kolom was eerst leeg. Als meer informatie nodig was om te kunnen bepalen wat de werkwoordstijd van een context was, bijvoorbeeld in het geval van een ambigu werkwoord, dan kan deze informatie wellicht verkregen worden uit de zin die in kolom O staat. Dit is de zin waar de context uit afkomstig is.



	B	C	D	E	F	G	H	I	J	K	L	M	N	O
264	19467	preterito perfecto simple	target	Pensé					VLfin					Pensé que , al cabo , era un domingo de menos , que mamá estaba ahora enterrada , que iba a
265	19348	preterito perfecto simple	target	trabajé					VLfin					Hoy trabajé duro en la oficina .
266	19437	preterito perfecto simple	target	estuvo					VEfin					El patrón estuvo amable .
267	19607	preterito perfecto simple	target	preguntó					VLfin					Me preguntó si no me encontraba demasiado cansado y quiso saber también la edad de mamá
268	19510	preterito perfecto simple	target	quiso					VLfin					Me preguntó si no me encontraba demasiado cansado y quiso saber también la edad de mamá
269	19672	preterito perfecto simple	target	Respondí					VLfin					Respondí que « unos sesenta años » , para no equivocarme e ignoro por qué pareció aliviado ,
270	19349	preterito perfecto compuesto	target	ha	servido				VHfin	VLadj				Por la tarde , disfruto menos porque la toalla giratoria que se utiliza está completamenta húme
271	19340	preterito perfecto simple	target	señalé					VLfin					Se lo señalé un día a mi patrón .
272	19631	preterito perfecto simple	target	Respondió					VLfin					Respondió que era , en efecto , deplorable , pero que se trataba , a fin de cuentas , de un detal
273	19403	preterito perfecto simple	target	quedamos					VLfin					La oficina da al mar y nos quedamos un momento mirando los cargueros en el puerto que ardí
274	19567	preterito perfecto simple	target	Llegó					VLfin					Llegó en ese momento un camión con estrépito de cadenas y explosiones .
275	19549	preterito perfecto simple	target	preguntó					VLfin					Emmanuel me preguntó « si nos subíamos » y eché a correr .
276	19692	preterito perfecto simple	target	siguió					VLfin					El camión siguió adelante y nos lanzamos en su persecución .
277	19478	preterito perfecto simple	target	apoyé					VLfin					Me apoyé el primero y subí de un brinco .
278	19341	preterito perfecto simple	target	subí					VLfin					Me apoyé el primero y subí de un brinco .
279	19508	preterito perfecto simple	target	ayudé					VLfin					Me apoyé el primero y subí de un brinco .
280	19523	preterito perfecto simple	target	preguntó					VLfin					Me preguntó « si estaba bien , a pesar de todo » .
281	19383	preterito perfecto simple	target	dije					VLfin					Le dije que sí y que tenía hambre .
282	19547	preterito perfecto simple	target	Comí					VLfin					Comí rápidamente y tomé café .
283	19698	preterito perfecto simple	target	tomé					VLfin					Comí rápidamente y tomé café .
284	19593	preterito perfecto simple	target	dormí					VLfin					Después volví a casa , dormí un poco porque había bebido demasiado vino y , al despertarme ,
285	19674	preterito perfecto simple	target	tuve					VLfin					Después volví a casa , dormí un poco porque había bebido demasiado vino y , al despertarme ,
286	19330	preterito perfecto simple	target	corrí					VLfin					Iba retrasado y corrí para subirme a un tranvía .
287	19455	preterito perfecto simple	target	Trabajé					VLfin					Trabajé toda la tarde .
288	19612	preterito perfecto simple	target	di	contra				VLfin	PREP				Al subir , en la oscura escalera , me di contra el viejo Salamano , mi vecino de piso .
289	19447	preterito perfecto compuesto	target	ha	llegado				VHfin	VLadj				A fuerza de vivir con él , solos los dos en una pequeña habitación , el viejo Salamano ha llegado
290	19678	preterito perfecto compuesto	target	ha	tomado				VHfin	VLadj				El perro , en cambio , ha tomado de su amo una especie de aire encorvado con el hocico hacia
291	19376	preterito perfecto compuesto	target	han	cambiado				VHfin	VLadj				Al cabo de ocho años , no han cambiado de itinerario .
292	19644	preterito perfecto compuesto	target	ha	olvidado				VHfin	VLadj				Cuando el perro ya se ha olvidado vuelve a tirar de su amo y vuelve éste a golpearlo y a insulta
293	19690	preterito perfecto simple	target	encontré					VLfin					Cuando lo encontré en la escalera , Salamano estaba insultando a su perro .
294	19642	preterito perfecto simple	target	Dije					VLfin					Dije : « Buenas tardes » , pero el viejo seguía con sus insultos .
295	19554	preterito perfecto simple	target	pregunté					VLfin					Le pregunté entonces qué le había hecho el perro .
296	19331	preterito perfecto simple	target	respondió					VLfin					No me respondió .

Afbeelding 8: Deel van Excel-bestand met Spaanse contexten

Aan de Nederlandse en Engelse contexten werd automatisch een werkwoordstijd toegekend. Deze moesten wel handmatig gecontroleerd worden. De reden hiervoor is terug te lezen onder het volgende kopje *Aandachtspunten uitvoering*. Het handmatig controleren van deze werkwoordstijden ging op dezelfde manier als het toekennen van de werkwoordstijden voor de Spaanse en Duitse contexten.

Resultaten verwerken

Als alle data geannoteerd is en deze met de juiste werkwoordstijd zijn getagd, is het tijd om dit te verwerken. De 5-tupels met de werkwoordstijden worden verwerkt zoals is beschreven onder het kopje *Methode*. Vervolgens zijn de semantische kaarten gemaakt met behulp van TimeMapping. Deze semantische kaarten zijn terug te vinden op de TimeAlign website.

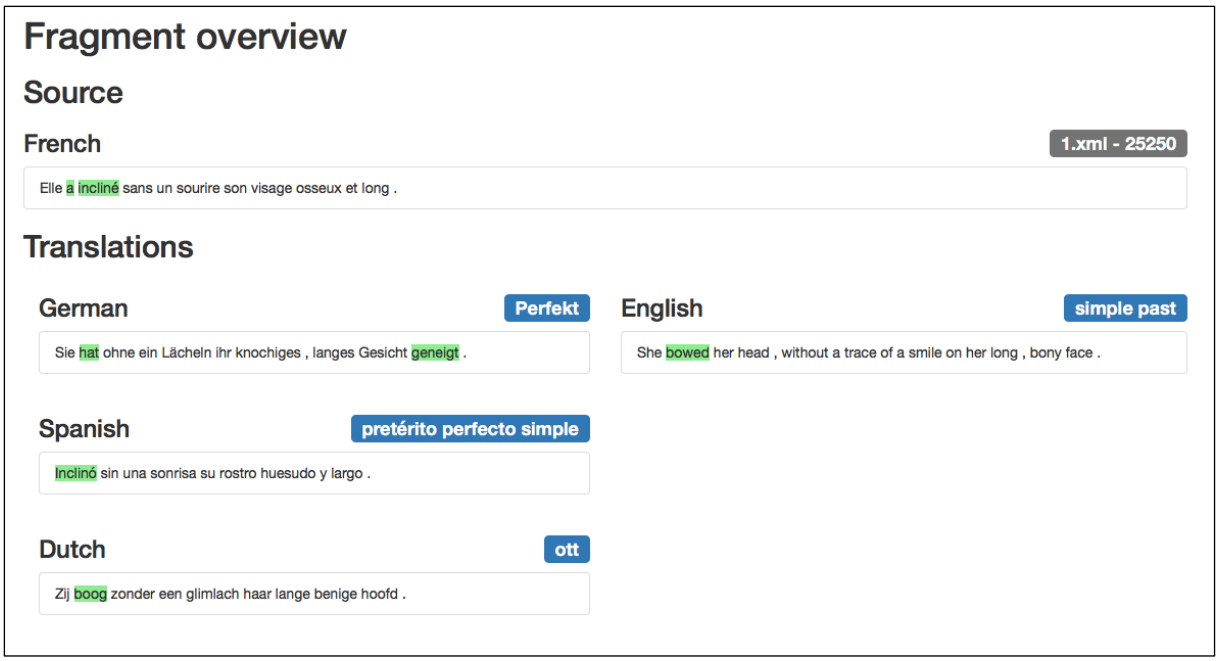
4.3 Aandachtspunten uitvoering

Tijdens de uitvoering van de nodige stappen om tot een resultaat te kunnen komen, zijn mij een aantal zaken opgevallen die het onderzoek hebben vertraagd.
Om eventuele vervolgonderzoeken voorspoediger te laten verlopen, zal het nuttig zijn om bewust te zijn van onderstaande punten.

Werkwoordstijden handmatig verbeteren

Nadat de semantische kaarten gevormd waren, werd al snel duidelijk dat er veel minder 5-tupels zichtbaar waren in de semantische kaarten, ongeveer 150 minder dan gehoopt. In eerdere onderzoeken zijn de contexten waarbij de optie is aangevinkt dat het geen correcte vertaling van de Perfect is, niet meegenomen bij het vormen van de semantische kaarten. De eerste gedachte was dus dat de vervallen contexten degene waren waarbij die optie was aangegeven. Na verder onderzoek hiernaar bleek dat bij het Nederlands achttien contexten waren aangevinkt met deze optie. Bij het Duits waren dit drie, bij het Engels zes en bij het Spaans zestien contexten. Totaal zijn dit 43 contexten, een stuk minder dan de verwachte 150 contexten.
Het bleek dus om iets anders te gaan. Het lag namelijk aan de automatische toekenning van de werkwoordstijden in het Nederlands en het Engels. Doordat ik tijdens het annoteren woorden heb geselecteerd die wel onderdeel uitmaakte van de vertaling, maar die niet verwerkt konden worden door het algoritme, zijn veel contexten geen werkwoordstijd toegekend. Enkele voorbeelden hiervan zijn ‘binnenreed’, ‘staken over’ en ‘nam afscheid’. Dit zijn respectievelijk de Onvoltooid Verleden Tijd van ‘binnenrijden’, ‘oversteken’ en ‘afscheid nemen’. Doordat deze vormen een onregelmatigheid bevatte, hierboven onderstreept aangegeven, konden deze niet herkend worden door het algoritme en werden deze de werkwoordstijd ‘other’ toegekend. Hierdoor zijn deze contexten niet meegenomen bij het vormen van de semantische kaarten, hoewel deze wel correct zijn. Het zou zonde zijn als deze correcte vormen niet meegenomen zouden worden bij het verwerken van de data, daarom was het nodig deze contexten handmatig een werkwoordstijd toe te kennen.

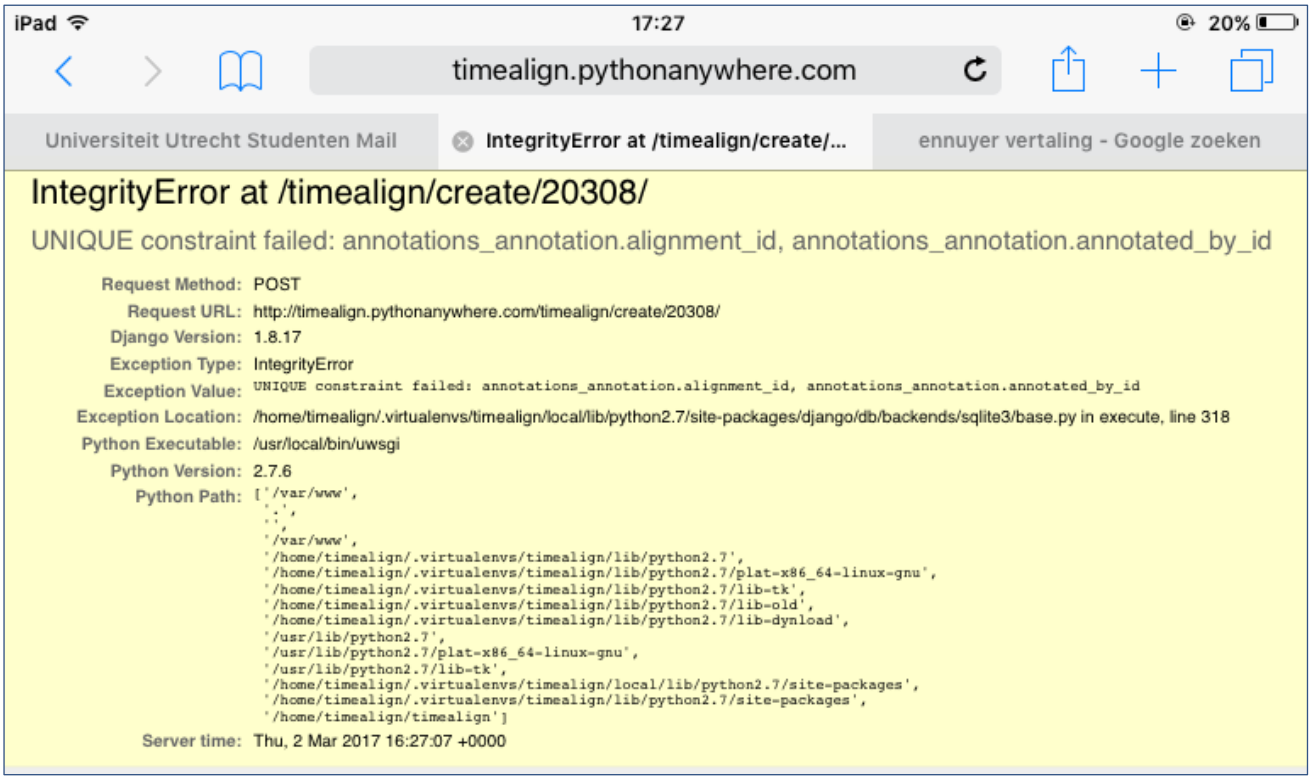
Ook werd duidelijk dat als er wel een werkwoordstijd automatisch was toegekend, deze niet altijd correct was. Bijvoorbeeld in het Nederlands waren een aantal contexten als een OTT getagd, die eigenlijk een OVT waren. Alle contexten die OTT waren toegekend, moesten dus per stuk nagegaan worden of deze toekenning wel correct was. Zie afbeelding 9 voor een voorbeeld waar de OTT foutief is toegekend.



Afbeelding 9: Voorbeeld foutief toegekende werkwoordstijd

Annoteren met een touchscreen

Als je annoteert op een apparaat met een touchscreen, is het nodig met iets meer beleid te werk te gaan. Er zit namelijk een klein verschil in welke handeling je moet uitvoeren voor het verwerken van een bepaalde context. Als de overeenkomende context in de vertaling uit één werkwoord bestaat, bijvoorbeeld '(zij) zag', en je selecteert 'zag', dan moet je tweemaal op de blauwe 'Submit' knop klikken. Als je een overeenkomende context in de vertaling selecteert die uit twee werkwoorden bestaat, bijvoorbeeld '(mijn baas) heeft gedacht', dan hoef je maar één keer op de 'Submit' knop te klikken. Op het moment dat je twee keer op de Submit knop klikt, terwijl je daarvoor al twee werkwoorden hebt geselecteerd, zal TimeAlign een foutmelding geven.
In afbeelding 10 is een voorbeeld te zien van zo'n foutmelding. Dit geeft aan dat de context al ingeleverd is en je deze niet opnieuw kunt behandelen. Als het ware heb je geprobeerd je annotatie twee keer achter elkaar in te leveren. Dit zorgt verder niet voor problemen wat betreft het verwerken van de werkwoorden, deze worden namelijk bij de eerste druk op de knop op de gangbare manier verwerkt. De tweede druk op de 'Submit' knop zorgt voor de foutmelding. Als de foutmelding in beeld komt, moet je terug gaan naar de vorige pagina. Dan kom je terug bij het fragment dat je als laatste hebt geannoteerd. Hierna kan je klikken op de oranje 'Go to another fragment' knop, om verder te gaan met het annoteren van de volgende context.



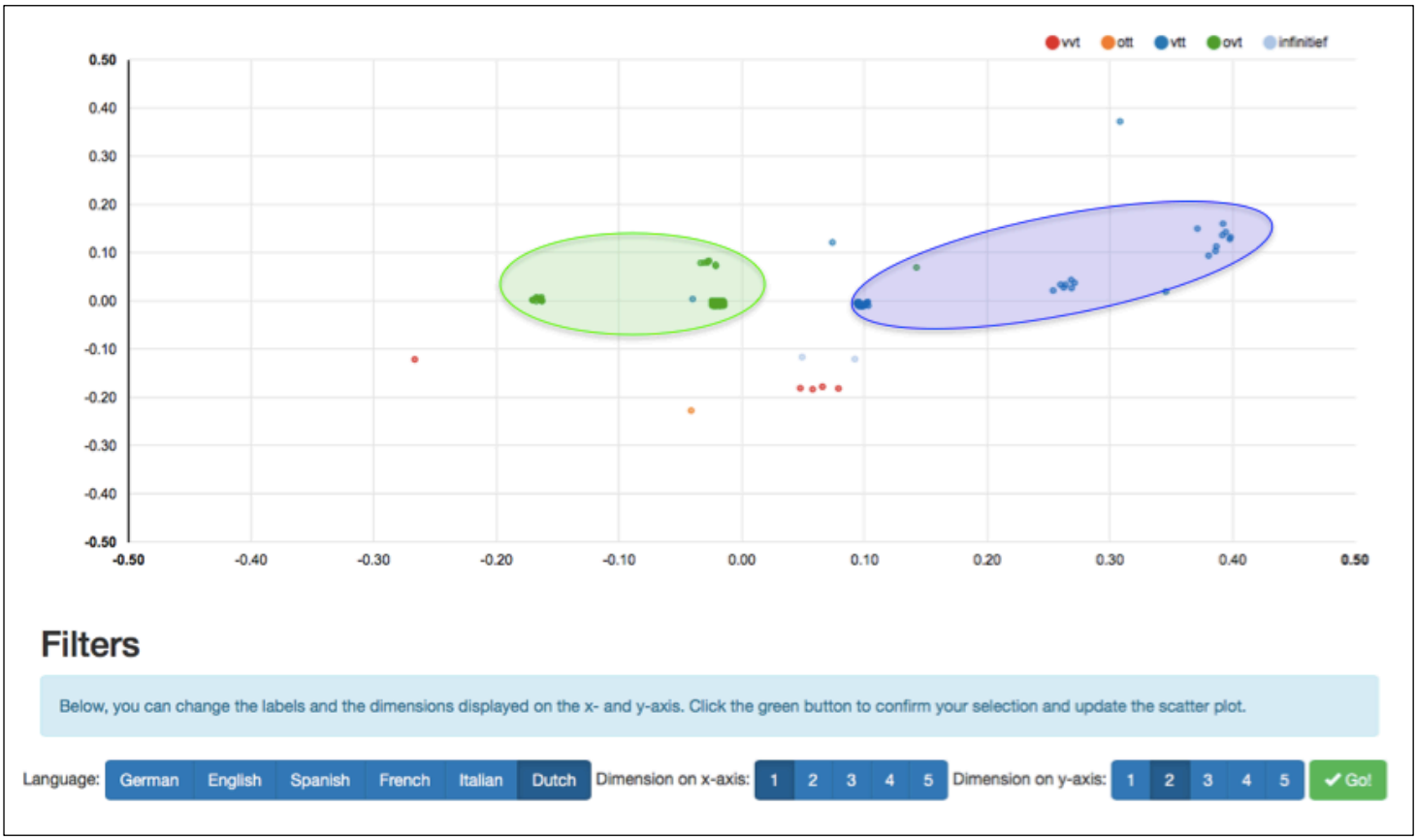
Afbeelding 10: Foutmelding tijdens annoteren met touchscreen

4.4 Resultaten

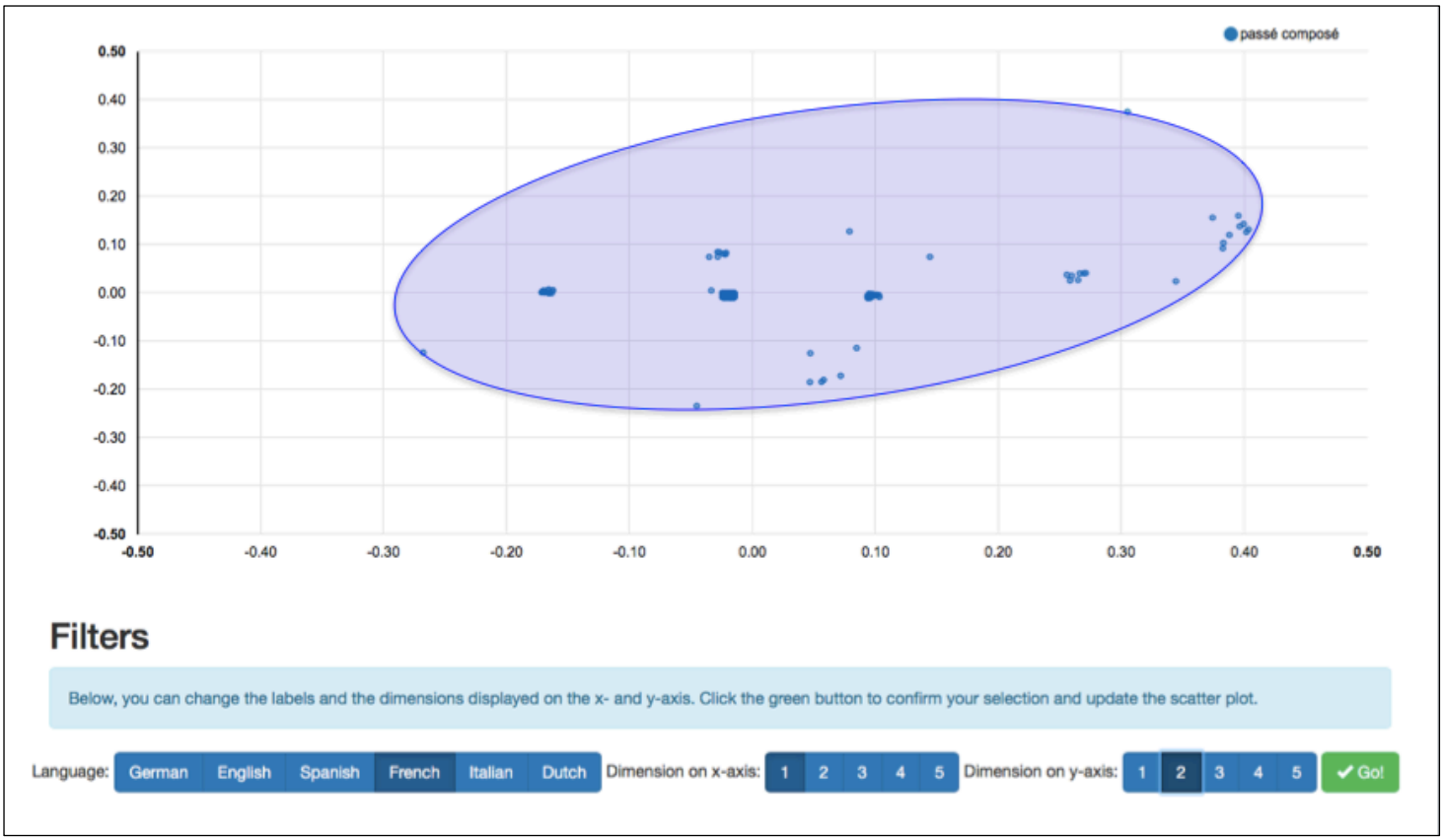
Nu alle data is verwerkt, kunnen de resultaten op een rijtje worden gezet. Met behulp van TimeMapping zijn de semantische kaarten gevormd, die zorgen voor een visuele representatie van de resultaten. Hieronder is per taal een semantische kaart te zien. Per taal zijn ook de aantallen van de gevonden werkwoordstijden in oplopende volgorde weergegeven in tabellen. De aantallen van de zeven meest voorkomende 5-tupels zijn in ook een tabel weergegeven, met daarna een voorbeeld van elke 5-tupel. Als laatste zijn de resultaten descriptief opgesteld.

Semantische kaarten

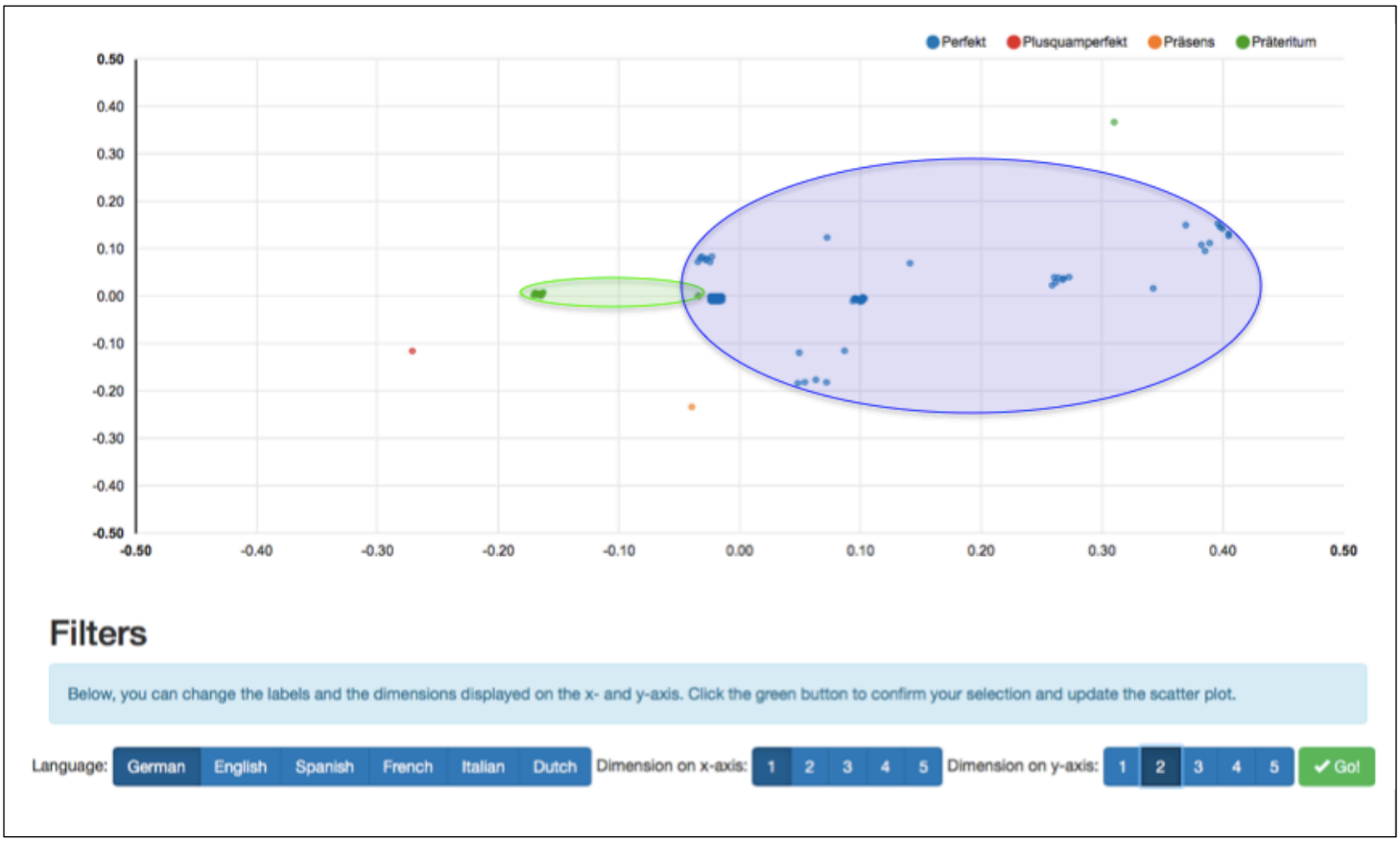
Zoals beschreven onder de kopjes *Methode* en *Uitvoering* zijn de semantische kaarten gevormd met het algoritme van TimeMapping. In afbeelding 11 tot en met 15 zijn de vijf kaarten van het Nederlands, Frans, Duits, Engels en Spaans te zien. Allen hebben op de x-as dimensie 1 en op de y-as dimensie 2. De clusters zijn aangegeven met gekleurde ovalen. De groene ovalen representeren de clusters van de OVT, Präteritum, Simple Past en Pretérito Perfecto Simple en de blauwe ovalen omvatten de VTT, Passé Composé, Perfekt, Present Perfect en Pretérito Perfecto Compuesto. De grootte van de clusters zegt niet één op één iets over de frequentie van de datapunten, want doordat overeenkomende 5-tupels afstand 0 hebben, verenigen zij zich als een klein groepje rond een punt. Hier kan de frequentie dus hoog zijn, maar de afstand 0, waardoor de spreiding van het cluster beperkt blijft.



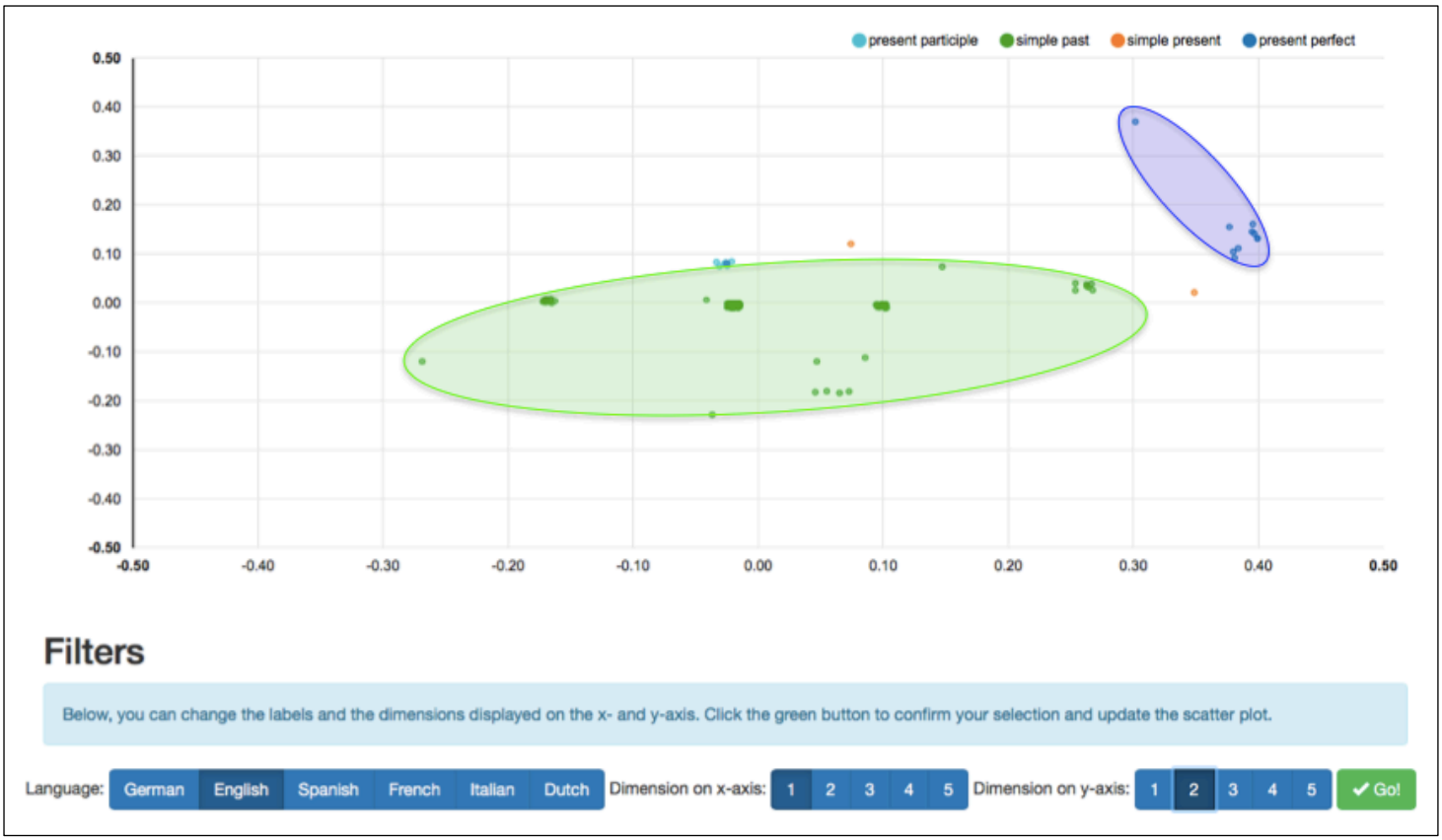
Afbeelding 11: Semantische kaart Nederlands



Afbeelding 12: Semantische kaart Frans



Afbeelding 13: Semantische kaart Duits



Afbeelding 14: Semantische kaart Engels

Afbeelding 15: Semantische kaart Spaans

Tabellen

In tabel 1 tot en met 4 zijn de aantallen van de gebruikte werkwoordstijden van het Nederlands, Duits, Engels en Spaans contexten in **oplopende volgorde** weergegeven. De tabel van het Frans is weggelaten, omdat alle 356 contexten de Passé Composé zijn. In tabel 5 zijn het aantal voorkomens van de zeven meest voorkomende 5-tupels in oplopende volgorde te zien. In afbeelding 16 tot en met 22 is van elk meest voorkomende 5-tupel een voorbeeld te zien.

Werkwoordstijd	Aantal
OVT	307
VTT	41
VVT	5
INF	2
OTT	1

Tabel 1: Werkwoordstijden Nederlandse contexten

Werkwoordstijd	Aantal
Perfekt	337
Präteritum	17
Plusquamperfekt	1
Präsens	1

Tabel 2: Werkwoordstijden Duitse contexten

Werkwoordstijd	Aantal
Simple Past	337
Present Perfect	11
Present Participle	6
Simple Present	2

Tabel 3: Werkwoordstijden Engelse contexten

Werkwoordstijd	Aantal
Pretérito perfecto simple	336
Pretérito perfecto compuesto	19
Pretérito imperfecto	1

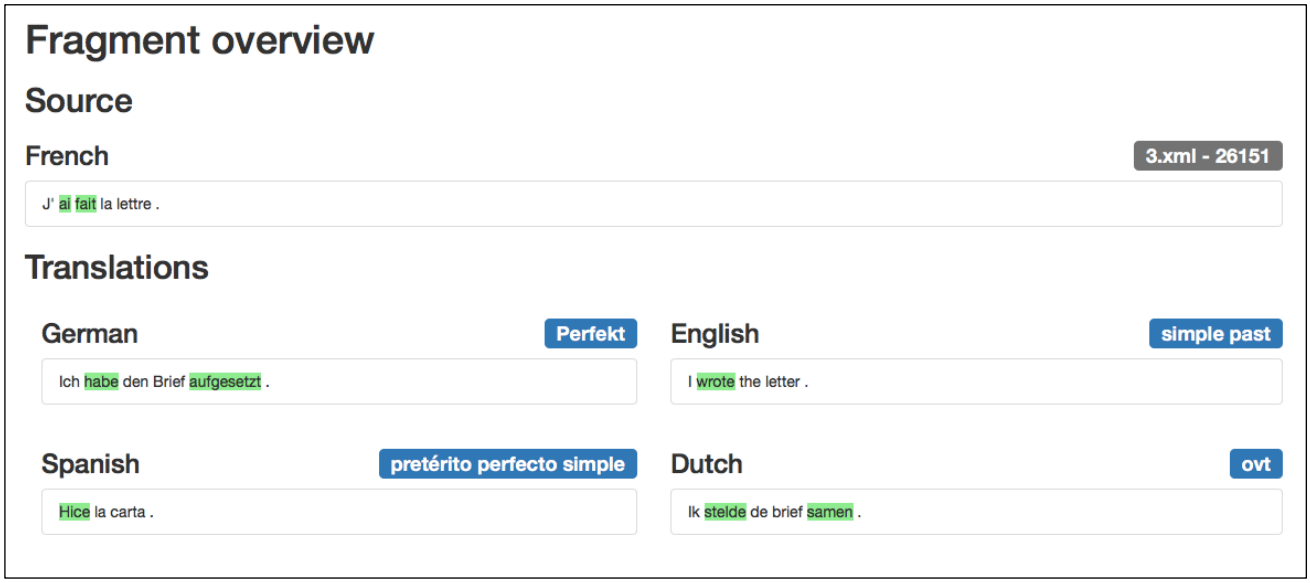
Tabel 4: Werkwoordstijden Spaanse contexten

5-tupel <NL, FR, DU, EN, SP>	Aantal
<OVT, Passé Composé, Perfekt, Simple Past, Pretérito Perfecto Simple>	284
<VTT, Passé Composé, Perfekt, Simple Past, Pretérito Perfecto Simple>	21
<OVT, Passé Composé, Präteritum, Simple Past, Pretérito Perfecto Simple>	15
<VTT, Passé Composé, Perfekt, Present Perfect, Pretérito Perfecto Compuesto>	9
<VTT, Passé Composé, Perfekt, Simple Past, Pretérito Perfecto Compuesto>	7
<OVT, Passé Composé, Perfekt, Present Participle, Pretérito Perfecto Simple>	6
<VVT, Passé Composé, Perfekt, Simple Past, Pretérito Perfecto Simple>	4

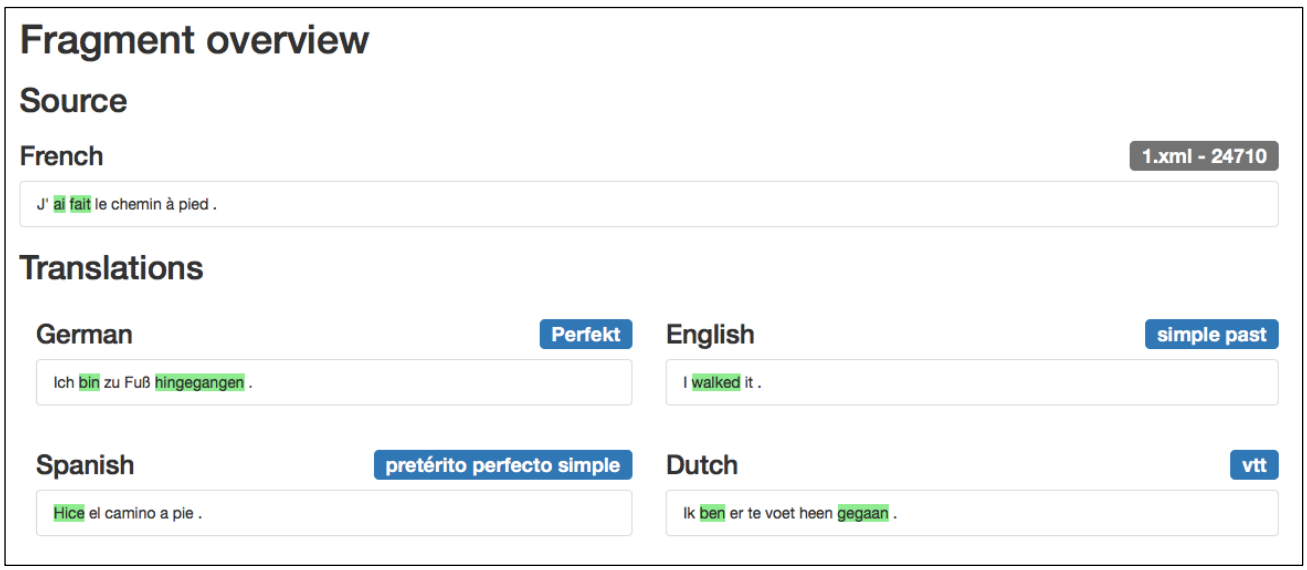
Tabel 5: Zeven meest voorkomende 5-tupels

Voorbeelden van elk 5-tupel

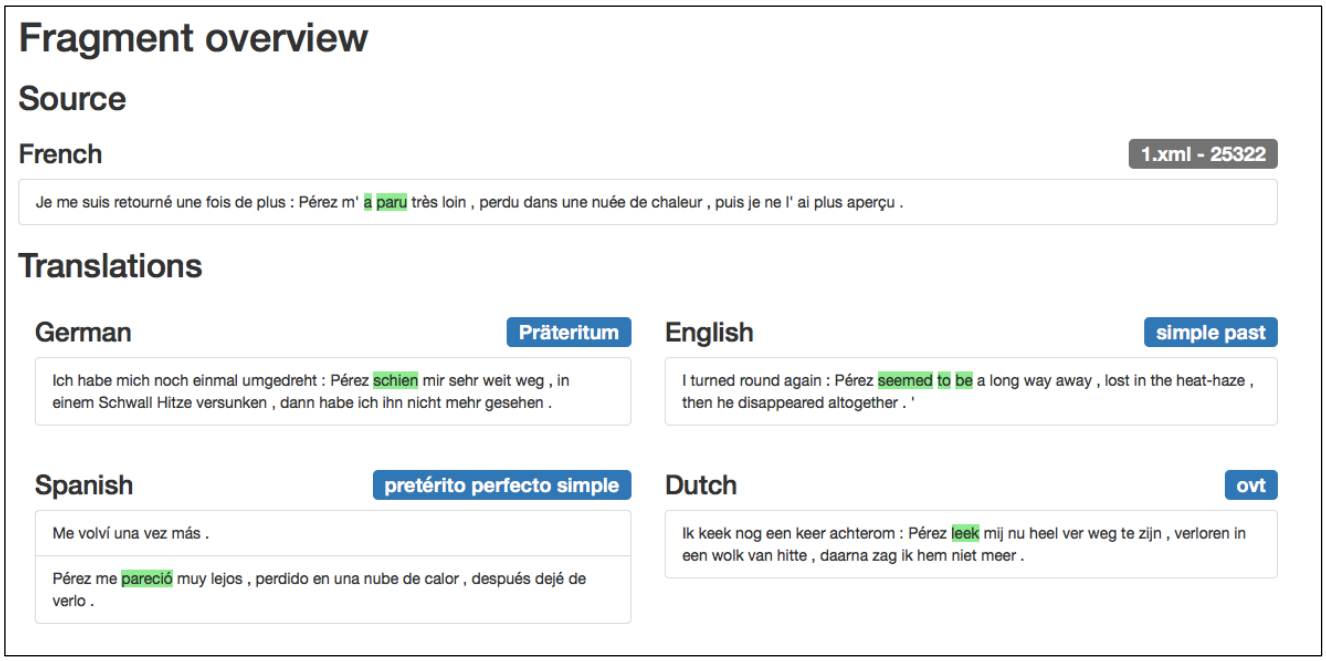
Hieronder is van elk soort 5-tupel uit bovenstaande tabel een voorbeeld te zien.



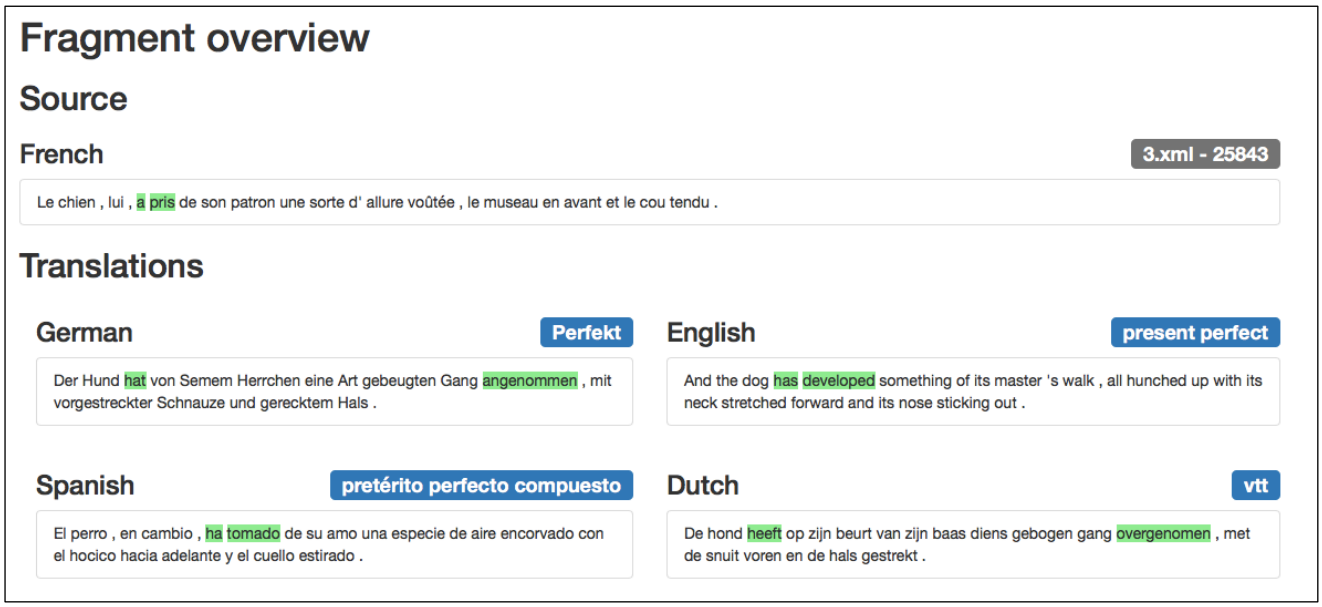
Afbeelding 16: 5-tupel <OVT, Passé Composé, Perfekt, Simple Past, Pretérito Perfecto Simple>



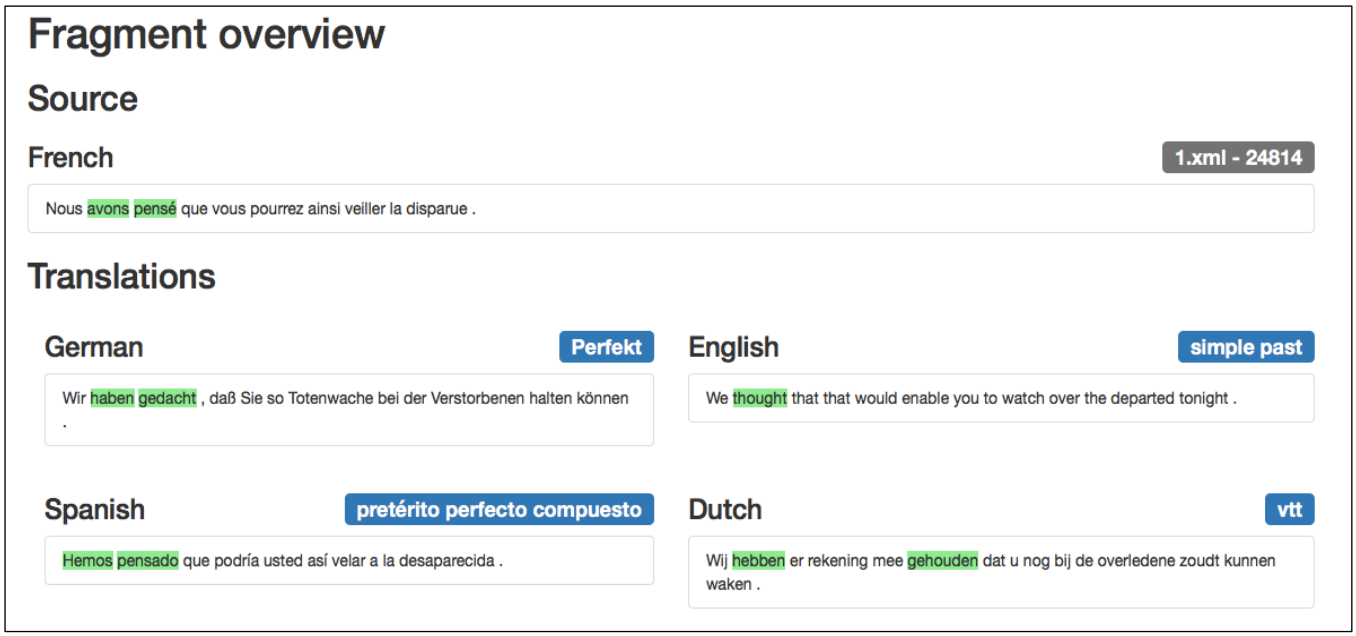
Afbeelding 17: Voorbeeld 5-tupel <VTT, Passé Composé, Perfekt, Simple Past, Pretérito Perfecto Simple>



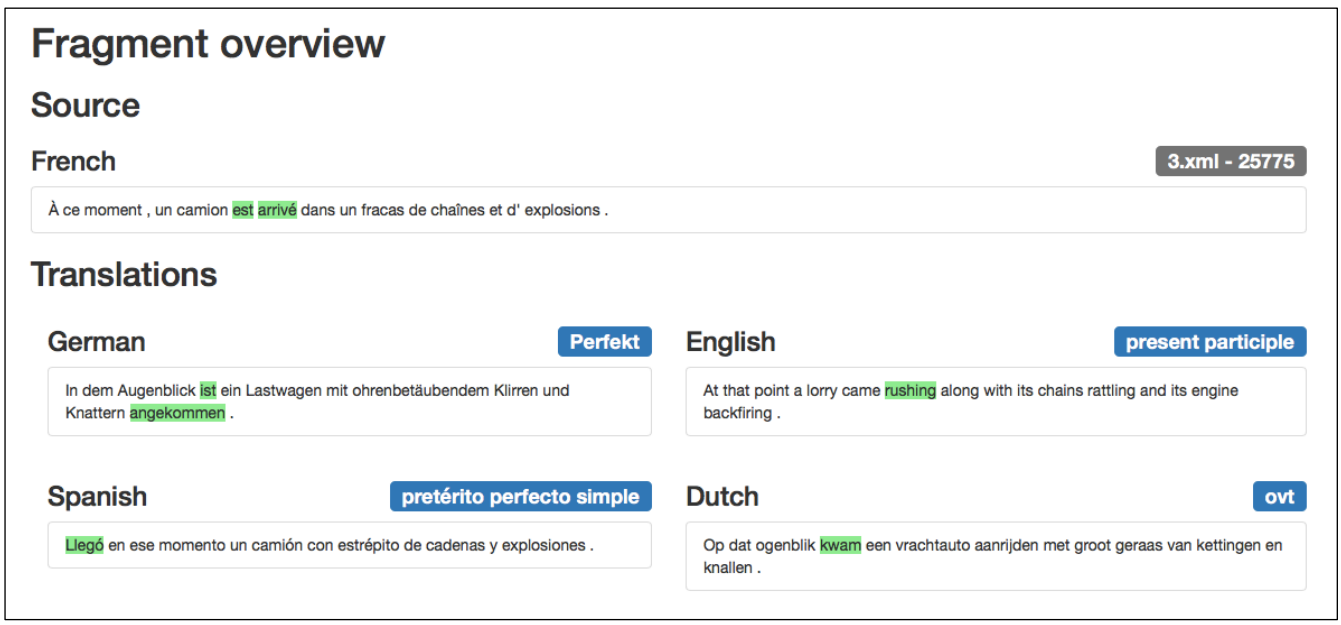
Afbeelding 18: Voorbeeld 5-tupel <OVT, Passé Composé, Präteritum, Simple Past, Pretérito Perfecto Simple>



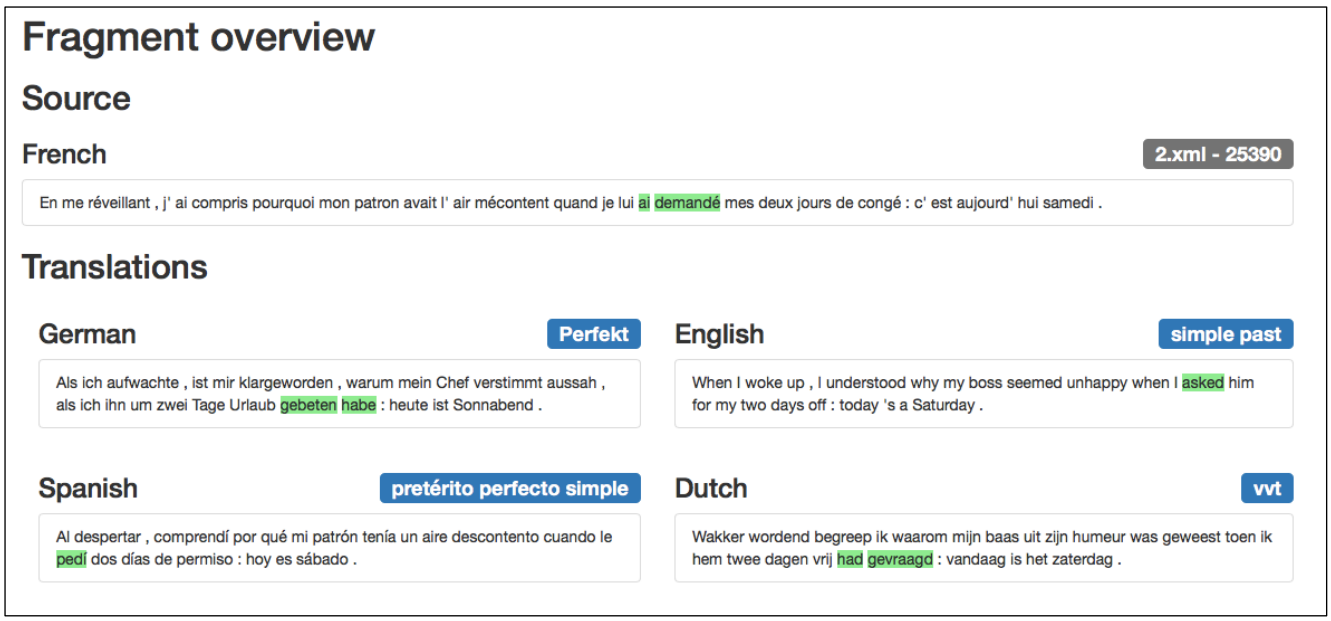
Afbeelding 19: Voorbeeld 5-tupel <VTT, Passé Composé, Perfekt, Present Perfect, Pretérito Perfecto Compuesto>



Afbeelding 20: Voorbeeld 5-tupel <VTT, Passé Composé, Perfekt, Simple Past, Pretérito Perfecto Compuesto>



Afbeelding 21: Voorbeeld 5-tupel <OVT, Passé Composé, Perfekt, Present Participle, Pretérito Perfecto Simple>



Afbeelding 22: Voorbeeld 5-tupel <VVT, Passé Composé, Perfekt, Simple Past, Pretérito Perfecto Simple>

Descriptief resultaat

Een duidelijk resultaat dat naar voren komt is dat de Perfect vanuit het Frans alleen in het Duits grotendeels met een Perfekt wordt vertaald. Hier staat tegenover dat deze in het Nederlands, Engels en Spaans voornamelijk wordt vertaald met respectievelijk een OVT, Simple Past en Pretérito Perfecto Simple. In het Nederlands wordt een deel met de VTT vertaald, namelijk 41 van de 356 contexten. In het Spaans (Pretérito Perfecto Compuesto) zijn dit 19 contexten en in het Engels (Present Perfect) 11 contexten.

In het Duits worden niet alle Perfects met een Perfekt vertaald, maar 17 contexten worden met een Präteritum vertaald.

De Perfect wordt nauwelijks met een Present vertaald. In het Nederlands en Duits is dit 1 context, in het Engels 2 en in het Spaans 0.

De tijden van de meest voorkomende 5-tupel is voor het Frans en Duits de Perfect, en voor het Nederlands, Engels en Spaans de Simple Past. Dit was ook te verwachten aan de hand van de gegevens per taal. De 5-tupel die daarna komt verschilt in één taal, het Nederlands is in deze 5-tupel namelijk ook de Perfect.

4.5 Conclusie

De eerste onderzoeksvraag is hoe de Perfect uit het Frans zich zal gedragen in de vertalingen. Gezien de duidelijke resultaten die uit dit onderzoek naar voren zijn gekomen, blijkt dat de Perfect vanuit het Frans zich over het algemeen als een Perfect gedraagt in het Duits, maar niet in de andere talen. In het Nederlands, Engels en Spaans gedraagt de Perfect zich meer als een Simple Past. Verder verdeelt de Perfect zich in het Nederlands ook gedeeltelijk over de Perfect. Dit zal te maken hebben met het gebruik van tijdsbijwoorden, zoals 'gister'. In het Spaans en Engels wordt een nog kleiner deel ook als een Perfect vertaald.
Dit komt helemaal overeen met de hypothese, waarin werd voorspeld dat de Perfect in het Duits een Perfect zou blijven, maar in het Nederlands en Engels niet. Bovendien werd verwacht dat in het Nederlands ook een deel met de VTT werd vertaald, wat ook het geval blijkt.
Over het Spaans is geen hypothese opgesteld, omdat hierover niet genoeg relevante informatie verkrijgbaar was. Nu is gebleken dat het Spaans zich hetzelfde gedraagt als het Nederlands en Engels, dus dit nemen we mee in de conclusie over de gedragingen van de Perfect.

De volgende onderzoeksvraag ging over eventuele clustering in de semantische kaarten. In het verlengde van de vorige onderzoeksvraag, had ik verwacht dat de clusters zouden ontstaan. Voor het Duits een cluster rond het Perfect, in de Nederlandse en Engelse kaarten rond de Simple Past. Deze voorspelling klopt ook. Er zijn hele duidelijk clusters te zien op de semantische kaarten van deze talen rond deze werkwoordstijden.
Tevens had ik verwacht dat in de Nederlandse semantische kaart een cluster te zien zou zijn rond de Perfect, doordat de aanwijzende temporele bijwoorden ervoor zouden zorgen dat deze in het Nederlands zo vertaald zou worden. Deze voorspelling is ook uitgekomen, want in de semantische kaart is een duidelijk blauw cluster te zien.
De vraag of er clustering zal onstaan in de semantische kaarten van mij onderzoek wordt beantwoord met een overtuigende 'ja!'.

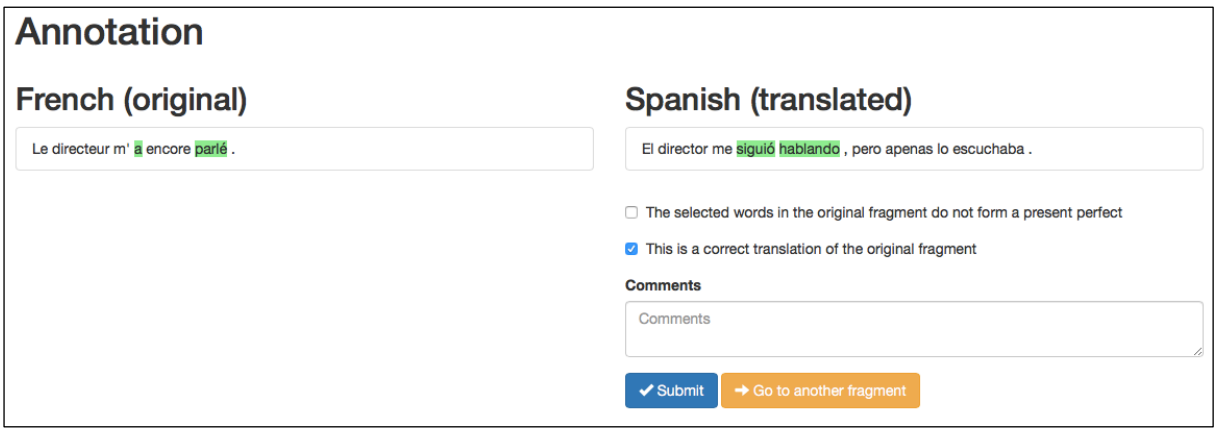
Aan het begin van het onderzoek was het nog onduidelijk hoe makkelijk OCR in gebruik zou zijn. Het was de vraag of het mogelijk zou zijn om lange stukken gedrukte tekst uit een boek te kunnen analyseren. Natuurlijk kun je een klein deel gedrukte tekst analyseren zonder dat het digitaal staat, maar voor dit onderzoek was het juist de bedoeling om veel contexten te verzamelen. Het is gebleken dat het gebruik van OCR voor weinig problemen zorgde en dus een zeer bruikbare optie is om gedrukte teksten digitaal te krijgen. Het antwoord op de vraag in hoeverre het mogelijk is om gedrukte teksten te analyseren, is dat dit zeker heel goed mogelijk is.

Als laatste was het de vraag of een verschil in de gedragingen van de Perfect merkbaar zou zijn tussen de gesproken taal en de geschreven taal. Mijn voorspelling hiervoor was dat ik meer variatie in het aantal werkwoordstijden verwachtte bij de gesproken taal. Dit omdat bij gesproken taal minder bewust wordt gekozen voor bepaalde woorden, terwijl dit bij geschreven taal heel bewust gaat. Deze voorspelling lijkt ook te kloppen. Bij eerder onderzoek in gesproken taal wordt gebruik gemaakt van meer soorten werkwoordstijden, namelijk tussen de 4 en 9 verschillende. Bij mijn onderzoek ligt dit aantal tussen de 3 en 5 werkwoordstijden. Er lijkt dus inderdaad meer variatie in werkwoordstijden bij gesproken taal.

4.6 Discussie

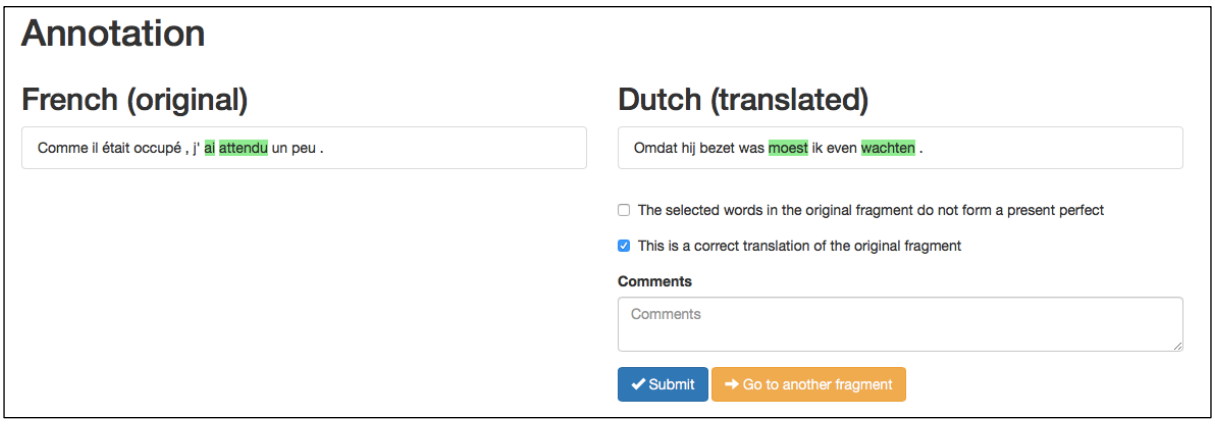
Keuze toekenning werkwoordstijd

Soms was het lastig te kiezen welke werkwoordstijd het beste paste. Bijvoorbeeld bij combinaties van werkwoorden. Een voorbeeld hiervan is te zien in afbeelding 23. Het werkwoord waar het om gaat is ‘parler’ en ‘hablar’. In de Spaanse vertaling is een woord toegevoegd, namelijk een vorm van ‘seguir’, wat zoiets als ‘voortgaan’ betekent. Dit deel van de Spaanse zin staat niet in de brontaal en zorgt ervoor dat de werkwoorden samen een continuïteit aangeven.
Voor dit onderzoek hebben we besloten om de werkwoordstijd van het hoofdwerkwoord, ‘siguió’ in dit geval, leidend te laten zijn. Dit komt doordat we de continuïteit nog niet nader gespecifieerd hebben en het nodig is beter na te denken hoe dit soort situaties benoemd moeten worden. In een vervolgonderzoek kan ervoor gekozen worden om dit te benoemen met een andere soort werkwoordstijd.



Afbeelding 23: Voorbeeld moeilijke keuze werkwoordstijd (FR-SP)

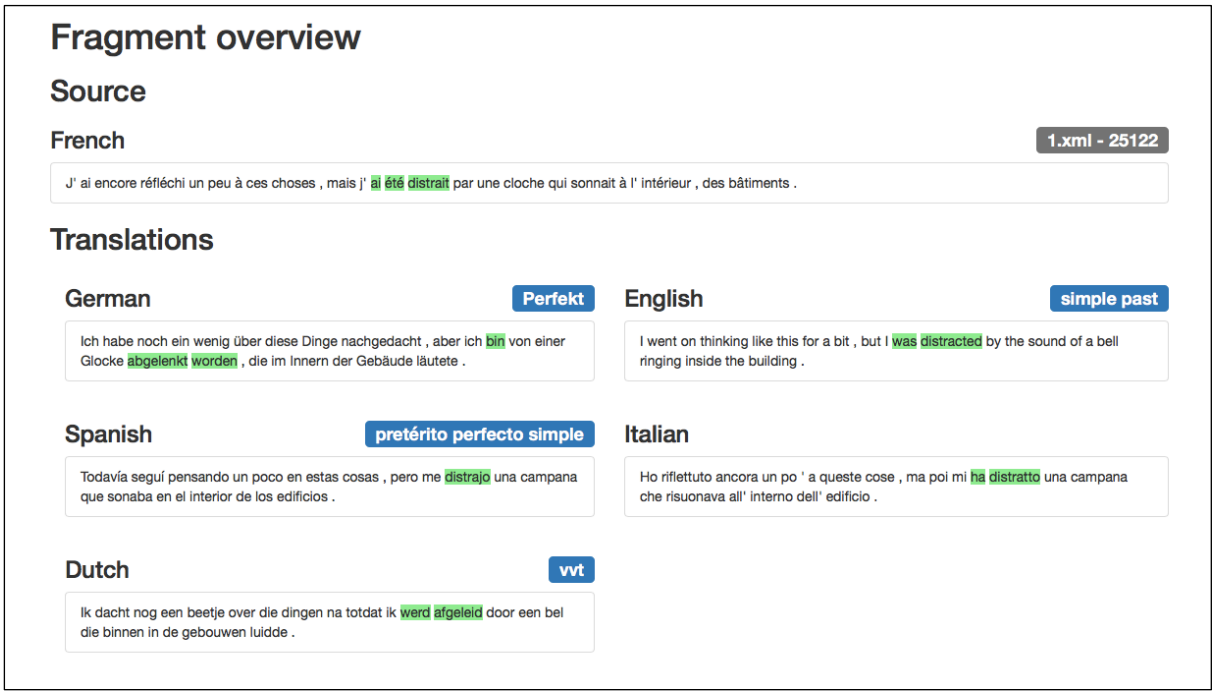
We hebben soortgelijke gevallen van combinaties van werkwoorden gezien in het Nederlands. Wederom is in de vertaling een extra werkwoord is toegevoegd, welke de functie van het hoofdwerkwoord draagt. Zie het voorbeeld in afbeelding 24.



Afbeelding 24: Voorbeeld moeilijke keuze werkwoordstijd (FR-NL)

Passieve vormen

Hebben de passieve contexten invloed op de resultaten? Dit is een vraag waar een antwoord op gevonden kan worden, als de passieve vormen apart behandeld worden.
Het zou kunnen dat het blijkt dat er minder Perfect vormen zijn dan dat de resultaten nu laten zien. Misschien is het mogelijk een algoritme te schrijven die de passieve vormen kan extraheren. Zie afbeelding 25 voor een voorbeeld van een passieve vorm, die als werkwoordstijd de VVT is toegekend, terwijl dit een OVT zou moeten zijn.
Dit is een voorbeeld voor het Nederlands, maar zulke gevallen komen voor in alle talen. Om deze reden zullen alle talen hierop gecontroleerd moeten worden. Zolang er geen algoritme voor geschreven is, is de data een stuk bewerkelijker. Hierdoor ben ik niet in staat geweest dit te verwerken binnen mijn onderzoek. Desondanks zou het interessant zijn om erachter te komen welke invloed de passieve vormen hebben op de resultaten.



Afbeelding 25: Voorbeeld passief werkwoord

Werkwoordstijden Duits

In het Duits wordt de Perfect grotendeels met de Perfekt vertaald. Toch wordt een niet verwaarloosbaar gedeelte met het Präteritum vertaald. In welke gevallen wordt deze tijd gebruikt en waarom? Dit is een vraag waar ik in dit onderzoek helaas geen antwoord op kan geven en waar bij vervolg onderzoeken aandacht aan besteed moet worden.
Kijkend naar de 17 contexten die in dit onderzoek met een Präteritum zijn vertaald, is het mij niet gelukt een duidelijk reden te ontdekken waarom de vertaler bij bepaalde contexten ervoor heeft gekozen om te vertalen met een Präteritum.

Verbetering semantische kaarten

Precies dezelfde 5-tupels hebben afstand 0 en liggen dus op dezelfde plek in de semantische kaart. Als je heel veel overeenkomende 5-tupels hebt, vormen deze een kluitje rond hetzelfde punt. Ik denk dat het mogelijk zou moeten zijn om de frequentie van de 5-tupels opvallender weer te geven. Bijvoorbeeld door één punt te maken van overeenkomende 5-tupels, die groter wordt naarmate er meer 5-tupels zich op dat punt bevinden. Als je op dat punt klikt, zou een genummerd scroll lijstje kunnen verschijnen waarin je een 5-tupel kunt selecteren. Als je een 5-tupel selecteert, word je op

dezelfde manier als nu doorgelinkt naar de vertalingen. Het scroll lijstje is genummerd zodat je gemakkelijker kunt onthouden welke 5-tupels je al eerder geselecteerd hebt. Als je terug klikt en weer door het lijstje gaat scrollen, kun je makkelijker een ander 5-tupel selecteren als je onthoudt welk nummer 5-tupel je al bekeken hebt. Dit is een idee voor het weergeven van de semantische kaarten, maar dit dient verder uitgewerkt te worden zodat het optimaal kan werken.
Ook zou er gewerkt kunnen worden met een heat map. Dit werkt vergelijkbaar met een kaart waarin temperaturen worden aangegeven. Hiermee zou een rode kleur kunnen aangeven dat de frequentie van de 5-tupels hoog is en een blauwe kleur geeft een lagere frequentie aan.

Verschil gesproken en geschreven taal

Ik kan niet met zekerheid zeggen of mijn verklaring, dat gesproken taal minder bewust wordt gekozen, ook klopt. Deze verklaring is nog niet af te leiden uit deze resultaten. Hiervoor zou meer onderzoek gedaan moeten worden binnen zowel de gesproken als de geschreven taal. Pas als meer data voorhanden is, kan geprobeerd worden een uitspraak hierover te doen en een verklaring te vinden.

Invloed tijdsbepalingen

Om nog dieper in te gaan op de data van dit onderzoek, kunnen de contexten die in het Nederlands zijn vertaald met een VTT onder de loep worden gelegd. Klopt de literatuur en wordt in het Nederlands inderdaad alleen met een VTT vertaald indien er een temporeel aanwijzend bijwoord in de zin staat? Of ontbreken deze bijwoorden soms ook? Om hier een duidelijke uitspraak over te kunnen doen, is het nodig alle contexten die als een VTT vertaald zijn te onderzoeken. Helaas paste dit niet binnen de tijd van mijn onderzoek, maar dit is zeker een onderdeel waar in een eventueel vervolgonderzoek naar gekeken kan worden.

4.7 Vervolgonderzoek

Verschillen tussen Engelse vertalingen

In mijn onderzoek heb ik gebruik gemaakt van de Engelse vertaling van Joseph Laredo (1982). Nu is in 2012 een nieuwe Engelse vertaling van Sandra Smith gepubliceerd. Enkele verschillen zijn op te merken tussen deze twee vertalingen. Een verschil wordt bijvoorbeeld in de eerste zin al duidelijk. Deze is namelijk in beide vertalingen op een andere manier vertaald. De eerste zin uit het orgineel: ‘Aujourd’hui, maman est morte.’ Dit is door Laredo vertaald als: ‘Mother died today’. Door Smith is dit op een persoonlijkere manier vertaald, namelijk door: ‘My mother died today.’ Dit is natuurlijk slechtst een klein verschil, maar deze manier van schrijven is in de gehele vertaling doorgezet. Mijn vraag voor een eventueel vervolgonderzoek is nu of deze andere manier van vertalen invloed heeft op de betekenis van de Perfects.

Talen ontbrekende perfect

Voor een vervolgonderzoek zou het interessant kunnen zijn om te kijken naar talen waarin geen Perfect bestaat. Talen zoals het Pools, Russisch en Chinees hebben geen Perfect. In welke vorm is de Perfect terug te vinden in deze talen? Mijn onderzoek kan gemakkelijk uitgebreid worden door de vertalingen in deze talen van de eerste drie hoofdstukken van L’Étranger te analyseren op dezelfde manier en deze resultaten samen te voegen en te vergelijken.

5 Logboek

Datum	Gedaan	Tijd
16 jan 2017	Artikelen Anonymous EACL (2017) en begin de Swart (2007) gelezen	5 uur
21 jan 2017	Spaanse vertaling overgetypt	3 uur
26 jan 2017	Nederlandse en Engelse vertalingen ingescand	1,5 uur
3 feb 2017	Engelse vertaling handmatig verbeterd	3 uur
5 feb 2017	Nederlandse vertaling handmatig verbeterd	3,5 uur
6 feb 2017	Spaanse vertaling ingescand + OCR	3 uur
9 feb 2017	Spaanse vertaling handmatig verbeterd	5 uur
11 feb 2017	Begin opzet document	3 uur
13 feb 2017	Methode uitgewerkt en artikel de Swart (2007) verder gelezen	6 uur
14 feb 2017	Bespreking met Henriëtte, Martijn en Bert	1 uur
18 feb 2017	Annoteren uitgeprobeerd	0,5 uur
25 feb 2017	Nederlandse contexten geannoteerd	3 uur
2 maart 2017	Vervolg Nederlandse contexten annoteren	1 uur
6 maart 2017	Engelse contexten geannoteerd	2 uur
6 maart 2017	Begin Spaanse contexten annoteren	0,25 uur
8 maart 2017	Vervolg Spaanse contexten annoteren	2,75 uur
14 maart 2017	Begin Duitse contexten annoteren	0,25 uur
17 maart 2017	Vervolg Duitse contexten annoteren	2,25 uur
21 maart 2017	Duitse werkwoordstijden handmatig toevoegen	3 uur
21 maart 2017	Begonnen met Spaanse werkwoordstijden toevoegen	0,5 uur
23 maart 2017	Artikel Nishiyama en Koenig gelezen	3 uur
29 maart 2017	Vervolg Spaanse werkwoordstijden handmatig toegevoegen	1,5 uur
30 maart 2017	Vervolg Spaanse werkwoordstijden handmatig toegevoegen	1 uur
30 maart 2017	Methode beschreven	5 uur
31 maart 2017	Methode en Uitvoering	8 uur
1 april 2017	Methode en Uitvoering	9 uur
2 april 2017	Methode en Uitvoering en semantische kaarten geanalyseerd	8 uur
3 april 2017	Nederlandse en Engelse werkwoordstijden handmatig verbeterd	10 uur
3 april 2017	Bespreking met Henriëtte en Bert	2 uur
4 april 2017	Inleiding en Literatuur	10 uur
5 april 2017	Inleiding en Literatuur	12 uur
6 april 2017	Methode en Uitvoering	9 uur
7 april 2017	Methode en Uitvoering	8 uur
9 april 2017	Aandachtspunten en op de TimeAlign website data aangepast	6 uur
10 april 2017	Alles tot Resultaten bijgewerkt en eerste concept opgestuurd	8 uur
11 april 2017	Resultaten	8 uur
12 april 2017	Methode, Uitvoering, Resultaten en Conclusie	8 uur
12 april 2017	Bespreking met Martijn	1 uur
13 april 2017	Inleiding, Methode en Referenties	6 uur
14 april 2017	Conclusie en Discussie	7 uur
15 april 2017	Methode	5 uur
17 april 2017	Gehele verslag gecontroleerd en opmaak aangepast. Conceptversie ingeleverd	9 uur
21 april 2017	Feedback verwerkt en eindversie ingeleverd	4 uur

6 Bibliografie

6.1 Primaire literatuur

Camus, A., (1942). *L'étranger* [De vreemdeling]. Parijs: Gallimard.

Camus, A., (1949). *De vreemdeling* (A. Morriën, Vert.). Epe: Hooiberg. (Orginele werk gepubliceerd 1942)

Camus, A., (2010). *Der Fremde* (U. Aumüller, Vert.). Reinbek: Rowohlt. (Orginele werk gepubliceerd 1942)

Camus, A., (1982). *The outsider* (J. Laredo, Vert.). London: Hamish Hamilton. (Orginele werk gepubliceerd 1942)

Camus, A., (2012). *El extranjero* (J. Á. Valente, Vert.). Madrid: Alianza Editorial. (Orginele werk gepubliceerd 1942)

6.2 Secundaire literatuur

Artikelen

Nishiyama, A., & Koenig, J. (2010). What is a Perfect State? *Language, 86*(3), 611-646.

De Swart, H. (2007). A cross-linguistic discourse analysis of the perfect. *Journal of pragmatics*, *39*(12), 2273-2307.

De Swart, H., & Molendijk, A. (2002). Le passé composé narratif: une analyse discursive de l'étranger de Camus. *Laca (2002)*, 193-211.

Van der Klis, M., Le Bruyn, B., & De Swart, H. (2017). Mapping the Perfect via Translation Mining. *EACL 2017*, 497.

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

Tiedemann, J. (2003). Recycling translations: Extraction of lexical data from parallel corpora and their application in natural language processing (Doctoral dissertation, Acta Universitatis Upsaliensis).

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V. & Nagy, V. (2005). Parallel corpora for medium density languages. *Proceedings of the RANLP 2005*, 590-596.

Schmid, H. (2013). Probabilistic part-of-speech tagging using decision trees. *New methods in language processing*, 154.

Verkleij, A. & Wimmers, V. (2016). Filmperspectief op de present perfect (Bacheloreindwerkstuk, Universiteit Utrecht)

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *LREC, 2012*, 2214-2218.

Lison, P., & Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation*.

Internetwebsites

Website TimeAlign onderzoek - timealign.pythonanywhere.com

Online OCR programma - www.onlineocr.net

PDF-bestanden splitten - www.splitpdf.com

Python-script treetagger-xml - github.com/mhkuu/treetagger-xml

Scikit-learn package - scikit-learn.org/stable/index.html

Visualisatie MDS - nvd3.org/

Spaanse bestelwebsite voor boeken – www.casadellibro.com

Grammatica en vervoegen werkwoorden - www.lingolia.com/en/

Verschil in Engelse vertaling - www.theguardian.com/books/2012/dec/09/outsider-albert-camus-smith-review