In recente jaren vondt er een verschuiving in het domein van *data-gedrevene machinevertaling* (MT) plaats: Er ontstond een grotere nadruk op de integratie van meer linguistisch gemotiveerde data voor de bouw van systemen die vloeiende vertalingen kunnen produceren. Een zeer nuttige bron is het parallelle treebank. Dat kunnen wij definiëren als een *parallel corpus* dat op verschillende niveaus gealigneerd is. Zo een bron is echter alleen op een zeer grote schaal bruikbaar, en het is dus onpraktisch om handmatig te produceren. Een door computers leesbaar parallel treebank om MT systemen op te trainen moet dus automatisch worden geproduceerd, terwijl men ook daarbij rekening moet houden om aan zekere nauwkeurigheidstandaarden trouw te blijven.

De automatische bouw van parallelle treebanks is wel een tamelijk recent verschijnsel, maar er werd reeds veel onderzoek op gedaan. Onze bijdrage vindt in de context van de ontwikkeling van een hybridisch syntactisch gebaseerd systeem voor automatische vertaling plaats. In dit werk experimenteren wij op en passen wij bestaande en nieuwe methoden van *tree alignment* toe om een verscheidenheid van grote parallelle treebanks in verschillende taalparen te produceren, maar met de focus op het taalpaar Nederlands-Engels. Ons belangrijkste doel is om de kwaliteit van het aligneren van *constituenten* te verbeteren.

Ten eerste bekijken wij het gebruik van zogenaamde *maximum entropy models* om het probleem van alignment binair te classificeren - dat wil zeggen, om voor elk paar knopen in de bomen te besluiten of ze gealigneerd moeten zijn of niet. Wij gebruiken maximum entropy models om systemen op een *discriminerende* manier te trainen door het gebruik van een log-lineair model: De gebruiker maakt een verzameling trainingsgegevens alsook een stel features aan, die de classifier kan gebruiken om te leren wanneer de paren knopen gealigneerd moeten worden. Gebaseerd op wat hij op de trainingsgegevens heeft geleerd, wordt het opgebouwde model op nieuwe data toegepast. Belangrijke features zijn word alignments, gelijkvormigheid van subbomen en de etiketten van knopen.

Wij onderzoeken welke factoren de kwaliteit van alignment beïnvloeden door middel van een meervoudige regressieanalyse en de berekening van correlaties. Ten tweede bekijken wij hoe wij regels kunnen gebruiken om de prestatie van de classifier te verbeteren. Vooral leggen wij klem erop om *meer* alignments te maken met het doel om *recall* te verhogen, omdat ons statistische model een hogere precisie maar een lagere recall heeft. Wij vinden dat heel simpele regels al positieve effecten op recall kunnen hebben. Wij experimenteren met verschillende handmatige regels alsook een heuristisch bottom-up algoritme. In het laatste geval gebruiken wij ook de relatieve gelijkvormigheid van de subbomen - waar de huidige paar kandidaatknopen als wortelknopen fungeren - als features.

Voor beide benaderingen vinden wij dat de kwaliteit van de word alignments zeer belangrijk is. De toepassing van de zogenaamde *well-formedness constraint* (WFC) heeft een positief effect op de precisie, maar leidt tot een verlaging in recall wanneer de word alignments afwijken. Wanneer wij regels gebruiken om

de WFC te verslappen, doordat wij word alignments toelaten die niet door beiden subbomen gedeeld worden, gaan niet alleen de recall maar ook de F-score omhoog.

Uiteindelijk passen wij wat wij uit deze experimenten hebben geleerd toe op een regelgebaseerd systeem dat de zogenaamde methode van *transformation-based learning* gebruikt. Een stel mogelijke regels worden automatisch aangemaakt door een combinatie van features die men handmatig opstelt. Door middel van trainingsgegevens wordt er telkens één beste regel gekozen, die dan aan een lijst wordt toegevoegd, waarop de regel op een ongealigneerde versie van de trainingsgegevens wordt toegepast. Wanneer men geen regel meer vindt die tot een verbetering leidt, stopt de training. De lijst regels past men dan toe op nieuwe data.

Wij vinden dat een tamelijk kleine maar goed gekoze set features genoeg is om hoge scores te bereiken. Zoals met onze regelgebaseerde aanpak laten wij enkele schendingen van de WFC toe voor een verhoogde recall. Behalve de gelijkvormigheid van subbomen bekijken wij ook andere structurele features zoals hoogte en unaire knopen. Niet alleen word alignments maar ook het type alignment - met meer of met minder vertrouwen - zijn belangrijke features. Ook gebruiken wij knoopetiketten in beperkte mate.

Wij demonstreren dat een combinatie van statistische en regelgebaseerde methoden, die een aantal benaderingen tot alignment alsook verschillende structurele beperkingen toepast, niet alleen de uiteindelijke kwaliteit van alignment verbetert, maar ook actief het zogenaamde *pijplijnprobleem* aanpakt. Als annotaties in een reeks van modulen streng sequentieel worden toegepast, zodat de ene module altijd voor de andere tewerk gaat, spreken wij van een pijplijn. Het pijplijnprobleem ontstaat als latere modules het werk van eerdere altijd overnemen en nooit mogen corrigeren. Het werk in dit proefchrift neemt een stap in de richting van een oplossing. Dat gebeurt doordat onze aanpak het aligneren van constituenten robuuster tegen fouten in het word alignment - hetgeen een direct en meetbaar effect op de kwaliteit van het aligneren van constituenten heeft - maakt, alsook relevante stijgingen in recall teweegbrengt.

Na het bespreken, demonstreren en het analyseren van de door ons gekozen systemen, passen wij onze output op het bovengenoemde syntactisch gebaseerde systeem toe en bekijken wij hoe zijn prestatie met een huidig state-of-the-art frase-gebaseerd statistisch systeem, dat op dezelfde datasets getrained is, vergelijkt. De resultaten laten zien dat het statistische systeem duidelijk beter is. Het syntactisch gebaseerde systeem levert echter soms vertalingen met een betere grammaticale kwaliteit. Ook laten wij zien dat vertaalmodellen die op hoge recall alignments - een combinatie van statistische (hoge precisie) en regelgebaseerde (hoge recall) alignments - getrained zijn, significant beter zijn dan de modellen die alleen op de door het statistische model geproduceerde hoge precisie alignments getrained zijn.

Ten slotte bespreken wij de resultaten, de wetenschappelijke betekenis van ons werk en wat nog in de toekomst ligt.