Cornetto:

Een lexicaal-semantische database voor taaltechnologie.

Het doel van Cornetto is het bouwen van een lexicale database voor het Nederlands met zowel semantische relaties als combinatorische informatie die nodig is om woorden te combineren. Semantische relaties vindt men o.a. in wordnets, waarin groepen synoniemen door middel van voornamelijk verticale subtype relaties met elkaar worden verbonden: *waakhond=>hond, koffie=>drank.* Daarnaast zijn er ook horizontale relaties: *waakhond=>bewaken, school=> lesgeven.*

Een netwerk van dergelijke conceptuele relaties vormt een krachtige kennisbron om te redeneren over teksten. Toch is taal niet alleen een kwestie van conceptuele relaties. Zo kunnen we op grond van het feit dat *koffie* en *thee* dranken zijn toch niet voorspellen dat we *koffie en thee zetten*, maar *limonade maken*. Hetzelfde geldt voor voorzetsels bij werkwoorden zoals *behandelen aan zijn verwondingen* maar *behandelen voor een ziekte.* Er is een heel scala aan dergelijke combinatorische informatie dat typisch is voor het Nederlands. Deze informatie moet gekoppeld moet worden aan de conceptuele informatie om het computers mogelijk te maken de betekenis van woorden in teksten te herkennen, maar ook om vloeiende teksten te genereren in toepassingen.

De methode die in Cornetto wordt gehanteerd is het samenvoegen van twee bestaande databases, namelijk het Nederlandse Wordnet (DWN) en het Referentie Bestand Nederlands (RBN) en die verder te verbeteren. Het DWN bevat verticale en gedeeltelijk ook semantische horizontale relaties. Het RBN bevat horizontale relaties en combinatorische informatie. Een van de eerste doelstellingen van Cornetto is van iedere woordbetekenis in DWN te bepalen met welke woordbetekenis in RBN die correspondeert en vice versa. Dit gebeurt door middel van een programma dat overlappende informatie uit beide bestanden vergelijkt. Vervolgens moet worden bepaald of de samengevoegde informatie semantisch valide en coherent is en in hoeverre uit de samengevoegde informatie verdere relaties kunnen worden afgeleid. Om dit te bereiken, wordt iedere woordbetekenis gekoppeld aan een formele ontologie (SUMO, DOLCE) die bepaalt welke relaties wel en niet mogelijk zijn. De relaties in het semantisch netwerk kunnen dan worden 'doorgerekend' om te zien of er inconsistenties optreden. Om het samenvoegen en vergelijken van de verschillende bronnen en stukken informatie mogelijk te maken, is er een speciale database ontwikkeld waarin alle informatie efficiënt aan elkaar gekoppeld is en d.m.v. een editor kan worden gecontroleerd en bewerkt. Database en editor worden op dit moment opgeleverd en het correctie-handwerk zal vanaf het najaar 2006 beginnen. Tegelijkertijd worden er automatische extractiemethodes ontwikkeld, die worden toegepast op een juridisch domein. De uiteindelijke data worden maart 2008 opgeleverd en komen beschikbaar via de TST centrale van de Nederlandse Taalunie.