Verschillende versies van het Eindhoven Corpus

Het Eindhoven Ccorpus – ook wel Corpus Uit den Boogaart (1975) genoemd – is de eerste verzameling van Nederlandstalige geschreven en (getranscribeerde) gesproken teksten die voor onderzoeksdoeleinden is gemaakt. Het corpus dateert uit de periode 1960-1973.

Het geschreven deel bevat tekstfragmenten van in totaal 600.000 woorden uit de periode 1964-1971. Het gesproken deel is aanzienlijk kleiner en bevat circa 120.000 woorden.

In 1989 is aan het Eindhoven-corpus het Renkema-corpus toegevoegd, een corpus met tekstfragmenten van correspondentie tussen de regering en de Staten-Generaal uit het parlementaire jaar 1975-1976.

Versiegeschiedenis

Het oorspronkelijke corpus is gebruikt als bron voor het in 1975 verschenen *Woordfrequenties in geschreven en gesproken Nederlands*. De versie 2.0.1 is aan de Vrije Universiteit Amsterdam (VU) tot stand gekomen. In de VU-versie is niet alleen het Renkema-corpus toegevoegd, ook is de weergave van de morfosyntactische codering gewijzigd en zijn er correcties uitgevoerd. Bij de ontwikkeling van de WOTAN 2-tagset heeft Hans van Halteren gedeeltes van het Eindhoven Corpus een upgrade gegeven.

Versie 2.5

Het INT heeft een nieuwe versie van het Eindhoven Corpus gemaakt, waarbij het corpus is omgezet naar TEI-XML en van gestructureerde metadata is voorzien. Daarnaast is de lemmatisering aangevuld en is er een reconstructie van het hoofdlettergebruik en de diakritische tekens gedaan met behulp van de Van Halterenversie en GiGaNT-Molex. De PoS-tagging (verrijking met woordsoort) is omgezet naar een met de CGN-tagset nauw verwante tagging, waarbij sommige kenmerken automatisch zijn toegevoegd, en daarna weer gedeeltelijk handmatig gecorrigeerd.