



Product Sheet COREA- coreferentiecorpus

Wat is het COREA-coreferentiecorpus?

Het COREA-coreferentiecorpus (COREA – Coreference Resolution for Extracting Answers) is een verzameling van Nederlandse teksten (ca. 150.000 tokens) waarin coreferentiële relaties systematisch gemarkeerd zijn.

Coreferentiële relatie	Toelichting	Voorbeeld
Identiteit of strikte coreferentie	antecedent en anafoor verwijzen naar exact hetzelfde object	[Xavier Malisse] ₁ heeft zich geplaatst voor de halve finale in Wimbledon. [Hij] ₁ zal dan tennissen tegen een onbekende tegenstander.
Deel-geheelcoreferentie	anafoor is een subdeel van het antecedent	Hij kon [zijn auto] ₁ niet meer starten. [De benzinetank] ₂ was leeg.
Type-tokencoreferentie	antecedent en anafoor hebben geen identiteitrelatie, maar beide refereren aan hetzelfde type semantisch object in de wereld	De man die [zijn rekening] ₁ aan zijn vrouw gaf was slimmer dan de man die [het] ₁ aan zijn minnares gaf.
Tijdsgebonden coreferentie	antecedent en anafoor refereren aan een object in de wereld op een specifiek tijdstip	[Bert Degraeve] ₁ , tot voor kort [gedelegeerd bestuurder] ₂ , gaat aan de slag als [chief financial and administration officer] ₃ .
Metonymie	naar het antecedent wordt verwezen door middel van beeldspraak	[Het paleis] ₁ is nooit veel meer, maar zeker nooit minder geweest dan [de exponent van de Belgische heersende klasse in haar conservatisme, in haar katholicisme, en met haar financieel-economische macht] ₂ .
Bezitsrelaties	relatie tussen bezit en bezitter	[Rita] ₁ sprak [[haar] ₁ tegenstander] ₂ ernstig toe.
Gebonden anaforen	anafoor verwijst naar een antecedent dat een algemene categorie uitdrukt	[Iedereen die iets nieuws wil bereiken] ₁ moet [zijn] ₁ nek uitsteken.
Predicatieve nominalen	predicatieve relaties bevatten informatie over het antecedent (geen coreferentiële relatie)	[Het mediabedrijf Vivendi Universal] ₁ is [een sterke stijger binnen de DJ Stoxx50] ₂ .
Apposities (repetitief, restrictief)	twee nominale constituenten beschrijven hetzelfde individu (deel kan weggelaten worden)	[Hu Jintao], [de president van China], hield een toespraak voor de VN. [De Nederlandse bankgroep] [ABN-AMRO] heeft over het tweede kwartaal van 2002 een nettowinst behaald van 534 miljoen euro.
Modaliteit en negatie	A is niet B (negatie) of A is B tot op zekere hoogte (modaliteit)	[Een partij als het CDA] is, volgens Bert de Vries en andere prominente partijleden, tegenwoordig [nou niet direct het toonbeeld van sociale betrokkenheid]. [De criminelen], [vaak genaturaliseerde Belgen],...

Tabel 1: coreferentiële relaties die handmatig gemarkeerd zijn in het corpus



Product Sheet COREA- coreferentiecorpus

Het COREA-coreferentiecorpus omvat:

- Krantenartikelen afkomstig uit het D-Coi-project (35.166 tokens)
- Getranscribeerde spraak uit het Corpus Gesproken Nederlands (CGN) (39.466 tokens)
- Lemma's uit de Spectrum (Winkler Prins) Medische Encyclopedie verzameld voor het IMIX ROLAQUAD project (74.445 tokens)

```
<COREF ID="3">Een 21-jarige dronkenlap</COREF> besloot maandagnacht <COREF ID="5005"
TYPE="IDENT" REF="3">zijn</COREF> roes uit te slapen op <COREF ID="9">de snelweg
A1</COREF> bij Naarden .
<COREF ID="12" TYPE="IDENT" REF="9">De politie</COREF> trof <COREF ID="14"
TYPE="IDENT" REF="5005">de man</COREF> slapend aan achter het stuur van <COREF
ID="5017" TYPE="IDENT" REF="14">zijn</COREF> auto , terwijl de motor nog draaide.
Volgens <COREF ID="22" TYPE="IDENT" REF="12">de politie</COREF> had <COREF
ID="23" TYPE="IDENT" REF="5017">hij</COREF> ruim twee keer zo veel gedronken als
toegestaan .
```

Figuur 1: output van de COREA-coreferentieservice

Software

Binnen het COREA-project is een systeem ontwikkeld dat automatisch coreferentiële relaties tussen nominale constituenten in teksten op kan lossen. Een demoversie van deze COREA-coreferentieservice is via de TST-Centrale online beschikbaar gesteld op www.inl.nl/producten, Productcatalogus, Tools, COREA-coreferentieservice.

Formaten

Het corpus is beschikbaar in twee formaten:

- MMAX-formaat: geschikt voor de annotatietool MMAX2 die gebruikt is voor het produceren van de annotaties
- XML-formaat: kan gevisualiseerd worden met behulp van een webbrowser zoals Firefox of Opera; de coreferentiële relaties worden automatisch gemarkeerd.



Product Sheet COREA- coreferentiecorpus

[Een markable] Een <i>niet-coreferentiële</i> markable	[Een markable] De geselecteerde markable
Het <i>semantische hoofd</i> is onderstreept	
[Een markable] Een <i>coreferentiële</i> markable	[Een markable] Een markable die <i>direct</i> naar de geselecteerde markable verwijst
[Een markable] Een <i>predicatief</i> verwijzende markable	[Een markable] Een markable die <i>indirect</i> naar de geselecteerde markable verwijst
[Een markable] Een <i>gebonden</i> verwijzende markable	[Een markable] Een markable waarnaar de geselecteerde markable <i>direct</i> verwijst
[Een markable] Een <i>bridging</i> verwijzende markable	[Een markable] Een markable waarnaar de geselecteerde markable <i>indirect</i> verwijst

p.1.s.1 [MADRID] -
p.1.s.2 [De Bulgaarse ex-koning Simeon II, wiens partij [de verkiezingen van [vorige maand]] won], meent dat [hij] [geen keus heeft dan [premier] te worden].
p.1.s.3 Tegen [de Spaanse krant El Pais] zei [hij] dat „[het volk] erom vraagt”.
p.1.s.4 [Simeon], die tussen [[1946] en [1996]] in [ballingschap] verbleef, won [de helft van [de zetels in [het Bulgaarse parlement]]].

Figuur 2: voorbeeld van een gemarkeerde corpustekst in xml

Distributie en prijzen

Het COREA-coreferentiecorpus is als downloadbaar bestand beschikbaar via de TST-Centrale (www.inl.nl/producten, Productcatalogus, Corpora, COREA-coreferentiecorpus).

prijs commercieel*: op aanvraag
prijs niet-commercieel*: € 0,-

* de prijzen zijn exclusief btw en eventuele verzend- en handlingkosten

Het COREA-coreferentiecorpus werd gefinancierd door Het STEVIN-programma van de Nederlandse Taalunie en is het resultaat van een samenwerking tussen de vakgroep Informatiekunde (Rijksuniversiteit Groningen), the CNTS Language Technology Group (Universiteit van Antwerpen) en Language and Computing (Sint-Denijs-Westrem). Alle rechten op het corpus zijn in handen van de Nederlandse Taalunie.

Colofon

Deze Product Sheet is een uitgave van de Centrale voor Taal- en Spraaktechnologie (TST-Centrale), een initiatief van de Nederlandse Taalunie (NTU) ondergebracht bij het Instituut voor Nederlandse Lexicologie (INL). De TST-Centrale wordt gefinancierd door de NTU.

E-mail: servicedesk@inl.nl

Ontwerp: Swantje Haage Ontwerp, Amsterdam