

CLARIN-NL BINNEN CLARIN

CLARIN is een Europees initiatief dat als doel heeft om e-science mogelijk te maken binnen de geesteswetenschappen en de sociale wetenschappen. Hiertoe brengt CLARIN instituten die taaltechnologieën en talige bronnen ontwikkelen samen met onderzoeksinstituten die bij hun onderzoek gebruik (zouden kunnen) maken van deze technologieën en bronnen.

Op 1 januari 2008 is de 3-jarige voorbereidende fase van CLARIN van start gegaan. Momenteel zijn ruim 30 instituten uit 25 verschillende landen als partner aangesloten bij CLARIN. Het Europese project is verantwoordelijk voor de coördinatie, samenwerking en het uitwisselen van gegevens tussen de verschillende deelnemende landen.

In ieder deelnemend land komt er een nationale organisatie die door de eigen overheid gefinancierd wordt. In Nederland is voor de komende 6 jaar ruim € 9 miljoen beschikbaar gesteld aan CLARIN-NL. De nationale projecten zijn verantwoordelijk voor het standaardiseren van bronmaterialen, het ontwikkelen en aanpassen van tools om met deze internationale standaarden om te gaan. CLARIN-NL zal bovendien de beoogde doelgroep opleiden en trainen in het gebruik van de infrastructuur ten bate van hun onderzoek.

Common Language Resources and Technology Infrastructure



ORGANISATIE EN CONTACT

Het uitvoerend bestuur van CLARIN-NL voert het dagelijks beleid uit, zoals dit is opgesteld door het bestuur. Het bestuur wordt hierbij geadviseerd door een nationaal en een internationaal adviespanel.

Uitvoerend bestuur:

Jan Odijk (Universiteit Utrecht)
Hans Bennis (Meertens Instituut - KNAW)
Arjan van Hessen (Universiteit Twente)
Peter Wittenburg (Max Planck Instituut, Nijmegen)

Voorzitter Bestuur:

Geert Booij (Universiteit Leiden)

Contact:

Postadres:
CLARIN-NL Programmabureau
Janskerkhof 13
3512 BL Utrecht

Bezoekadres:
Achter de Dom 24
3512 JP Utrecht

Telefoon: 030-253 6730
Fax: 030-253 6000
Website: www.clarin.nl
E-mail: clarinnl@uu.nl

WWW.CLARIN.NL

Common Language Resources and Technology Infrastructure



De ontwikkeling van een Europese digitale infrastructuur waarin talige hulpbronnen en technologieën voor alle onderzoekers eenvoudig beschikbaar zijn.

CLARIN-NL is in 2009 van start gegaan binnen de overkoepelende organisatie CLARIN-EU.

WWW.CLARIN.NL



MISSIE CLARIN-NL

In de afgelopen decennia zijn veel talige resources ontwikkeld (zowel technologieën als data). Deze resources worden tot nu toe vooral gebruikt door onderzoekers uit de taal- en spraaktechnologie. Onderzoekers uit andere vakgebieden hebben slecht toegang tot deze nieuwe ontwikkelingen en beseffen vaak niet wat er allemaal mogelijk is met deze resources.

CLARIN richt zich daarom nadrukkelijk op wetenschappers uit de humaniora en de sociale wetenschappen die in hun onderzoek gebruik maken van grote hoeveelheden teksten. CLARIN-NL beoogt het mogelijk te maken hen eenvoudig toegang te geven tot de relevante materialen en hen inzicht te geven in de vele mogelijkheden die de verschillende bestaande talige hulpmiddelen en instrumenten kunnen bieden voor hun onderzoek. Het CLARIN-NL programma, dat een looptijd heeft van 6 jaar, is geslaagd als na afloop grote groepen niet-technische onderzoekers in hun onderzoek gebruik maken van de gecreëerde infrastructuur.

Hiernaast vindt u twee showcases om de functionaliteiten van CLARIN te demonstreren.



SHOWCASE 1

Een historicus in Amsterdam wil vergelijken hoe Nederlandse en Duitse kranten schrijven aan de vooravond van de Tweede Wereldoorlog. Zowel in Nederland als in Duitsland zijn de meeste relevante kranten uit die periode gedigitaliseerd. Echter, kranten die nu nog bestaan hebben dat elk op hun eigen wijze gedaan, waarbij het er vooral om ging de kranten op datum te kunnen terugvinden. De niet meer bestaande krantentitels werden door nationale bibliotheken gedigitaliseerd, waarbij vooral aandacht werd besteed aan het toevoegen van zoveel mogelijk metadata. Een bevriende collega van de universiteit van Ulm heeft een mooi programma geschreven waarmee do-

cumenten taal-onafhankelijk geïndexiseerd kunnen worden. Het probleem is dat het programma verwacht dat de data volgens het formaat van de Deutsche Digitale Bibliothek beschreven zijn. Om het onderzoek met zoveel mogelijk kranten te kunnen doen, moeten al deze kranten dus op dezelfde manier benaderbaar zijn.

Als beide bibliotheken hun krantendatabase herschrijven (=cureren) naar een CLARIN-compatible formaat en het classificatieprogramma een speciale schil krijgt om met CLARIN-compatible data te werken, kan de Amsterdamse onderzoeker alle bronnen in één systeem raadplegen.

SHOWCASE 2

Een taalkundige in Utrecht werkt aan een artikel over flectie in het Nederlands. Om goed te begrijpen hoe het verschijnsel van agreement werkt, is het nodig dat zij toegang heeft tot flectievervalslijsten in variëteiten van het Nederlands. Niet alleen ten aanzien van buigingsvervalslijsten in hedendaags standaard Nederlands, maar ook van de verscheidenheid in oudere stadia, dialecten en sociaal bepaalde variëteiten. Bovendien zou het wenselijk zijn om te kijken naar wat er aan flectievervalslijsten optreedt in eerste en tweede taalvervalsing. Om dit te kunnen doen moet deze

taalkundige toegang hebben tot een groot aantal verschillende resources die elk hun eigen specifieke eigenschappen hebben, en vaak niet of nauwelijks toegankelijk zijn voor een buitenstaander. Als CLARIN eenmaal in werking is dan kan zij langs een vergelijkbare weg toegang krijgen en zoeken in databases als Corpus Gesproken Nederlands aan het INL, dialectdatabases als de Morfologische en Syntactische Atlas van de Nederlandse Dialecten van het Meertens Instituut, historische databases zoals het Corpus Gijsseling, en taalvervalsingdatabases als Childes. Uit verschillende datavervalsingen stelt zij op vrij eenvoudige wijze haar eigen flectiecorpus samen dat een basis vormt voor nieuwe inzichten. Zo is zij in staat een nieuwe theorie te ontwikkelen over het proces van deflectie in het Nederlands, op grond van een tot dan toe ongekend rijke en gedetailleerde verzameling hedendaagse en vroegere vervalslijsten.

