Informatieblad

Frequentielijsten Corpora

Frequentielijsten Corpora

Wat is het product Frequentielijsten Corpora?

Het product Frequentielijsten Corpora is een verzameling lijsten van de 5000 meest voorkomende woorden en hun frequentie in een aantal corpora die beschikbaar zijn bij de TST-Centrale.

Welke corpora werden gebruikt?

- Het **5 Miljoen Woorden Corpus 1994** bestaat uit een verzameling teksten van o.a. boeken, tijdschriften, kranten en tv-nieuwsuitzendingen, die dateren uit de periode 1970-1994.
- Het **27 Miljoen Woorden Krantencorpus 1995** is een verzameling krantenteksten uit de NRC. De teksten verschenen in 1994 of 1995.
- Het **38 Miljoen Woorden Corpus 1996** bestaat voornamelijk uit krantenteksten en teksten uit boeken, tijdschriften, troonredes en nieuwsuitzendingen, afkomstig uit de periode 1982-1995. Bovendien bevat het corpus ook nog een verzameling juridische teksten die dateren uit de periode 1814-1989.
- Het **PAROLE-corpus 2004** is een verzameling modern-Nederlandse teksten die voornamelijk afkomstig zijn uit kranten, tijdschriften en boeken uit de periode 1982 tot 1998. Het corpus bevat ruim 20.000.000 woorden.
- Het **Corpus Gesproken Nederlands** is een verzameling van ongeveer 900 uur gesproken Standaardnederlands afkomstig van Vlamingen en Nederlanders. Het bevat ongeveer 9.000.000 woorden en werd onderverdeeld in vijftien categorieën op basis van de parameters voorbereid – spontaan, uitgezonden – niet-uitgezonden, monoloog – dialoog/multiloog,…
- Het **Algemeen Nederlands Woordenboekcorpus** bestaat voornamelijk uit internetteksten en bevat ook krantenmateriaal en literaire werken zoals essays, romans, verhalen,… Het corpus is nog in opbouw en zal bij afronding ongeveer 104 miljoen woorden bevatten.
- Het **Eindhoven Corpus** is een verzameling Nederlandstalige geschreven en gesproken teksten uit de periode van 1960 tot 1973. Het corpus bevat ongeveer 720.000 woorden.
- Het **D-Coi-corpus** is een tekstverzameling hedendaags geschreven Nederlands van ongeveer 54 miljoen woorden. De teksten zijn afkomstig van o.a. (online) nieuwsbrieven, websites, (online) magazines, brochures, juridische teksten, handleidingen enz.
- **SONAR-500** bevat meer dan 500 miljoen woorden tekst afkomstig uit uiteenlopende domeinen en genres. Alle teksten werden getokeniseerd, ge-POS-tagd en gelemmatiseerd. Ook de named entities werden gelabeld. Alle annotaties van SoNaR-500 werden automatisch geproduceerd. De frequentielijst omvat de nieuwemediateksten (tweets, chats en sms'en), die ook verzameld werden in het kader van het STEVIN-project SoNaR.

Informatieblad

Frequentielijsten Corpora

Welke lijsten werden er opgesteld?

Van het Eindhoven Corpus is er enkel een frequentielijst van de woordvormen beschikbaar en voor de andere corpora werd er telkens een woordvorm- en een lemmafrequentielijst aangemaakt.

Opgelet

Bij het onderling vergelijken van de frequentielijsten van de hierboven beschreven corpora, moet rekening gehouden worden met de volgende factoren :

- de corpora verschillen enorm in grootte (vb. ANW-corpus = 104.000.000 woorden vs. Eindhoven Corpus = 720.000 woorden).
- het gaat om verschillende soorten corpora. Zo bevat het Corpus Gesproken Nederlands bijna uitsluitend gesproken taal, terwijl er in het 27 Miljoen Woorden Krantencorpus 1995 enkel geschreven taal opgenomen werd. En het is ook evident dat het taalgebruik in juridische teksten (vb. in het 38 Miljoen Woorden Corpus 1996) afwijkt van het taalgebruik in krantenartikels (vb. in het 5 Miljoen Woorden Corpus 1994).
- de corpora bevatten materiaal uit verschillende periodes, wat een invloed heeft op de gebruikte woordenschat. Zo werden er in het 38 Miljoen Woorden Corpus 1996 teksten opgenomen uit de 19e eeuw, terwijl alle artikels uit het 27 Miljoen Woorden Krantencorpus 1995 gepubliceerd werden in 1994 of 1995.