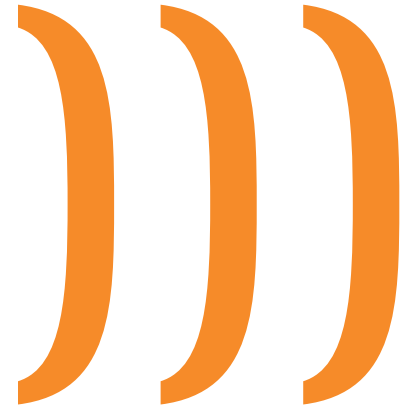


TST-nieuws

2007 nr 2



In deze tweede nieuwsbrief van 2007 wordt aandacht besteed aan de eerstvolgende themadag gericht op studenten, en aan één van de nieuwste producten van de TST-centrale; de Memory-Based Tagger-Lemmatizer. Meer nieuws is te vinden op de website www.tst.inl.nl. Sinds kort zijn ook alle voorgaande nieuwsbrieven hier online te bekijken. Reacties naar aanleiding van de nieuwsbrief kunt u sturen naar tst-info@inl.nl.

Product Sheets

Vier nieuwe Product Sheets van beschikbare TST-producten zijn op de website geplaatst: het 5 Miljoen Woorden Corpus 1994, het 27 Miljoen Woorden Krantencorpus 1995, het 38 Miljoen Woorden Corpus 1996 en e-Lex 1.1. De Product Sheets (PDF) geven zowel inhoudelijke als technische informatie van het product zo beknopt en duidelijk mogelijk weer. Zie www.tst.inl.nl, Producten, en klik op het icoontje in de meest rechtse kolom.

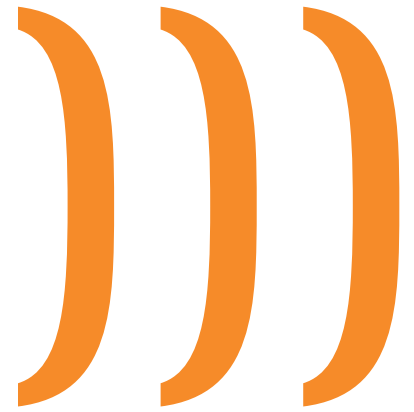
Stagedag studenten taal en spraak

De halfjaarlijkse themadag van de TST-centrale is dit keer speciaal toegespitst op studenten. Het is een stagedag waarop studenten kennis kunnen maken met de taal- en spraaktechnologie in het bedrijfsleven, en met de stagemogelijkheden op dat gebied.

Op 23 mei aanstaande komen stagiairs en medewerkers van bedrijven op het gebied van taal- en spraaktechnologie vertellen over hun ervaringen in de TST; wat doet bijvoorbeeld een taalkundige bij een taaltechnologiebedrijf of hoe verloopt een stageproject op het gebied van automatische e-mailbeantwoording?

Themadag
“Stagedag voor studenten taal en spraak”
Wanneer: 23 mei 2007
Waar: Hogeschool Domstad Utrecht
Hoe laat: van 11.00 – 17.00
Toegang: gratis

Tussen de presentaties door is er gelegenheid om de informatiemarkt te bezoeken, waar stands staan van bedrijven zoals Telecats, Polderland, Q-go en Dutcheer. Op de informatiemarkt zijn medewerkers van de bedrijven beschikbaar voor vragen en er worden promotiematerialen en vacante stageplaatsen aangeboden. Voor inschrijving en meer informatie over het programma en de locatie, zie www.tst.inl.nl



Uitgelicht: Memory-based Dutch Tagger-Lemmatizer

Voor het taggen en lemmatiseren van – gesproken of geschreven – Nederlandse teksten, heeft de TST-centrale sinds kort een nieuw product beschikbaar: de MBDDagLem.

De tagger-lemmatizer is ontwikkeld aan de Universiteit van Tilburg (ILK Research Group) en gebaseerd op memory-based learning technieken. De tagset die gebruikt wordt, is ontwikkeld in het kader van het Corpus Gesproken Nederlands. Deze set is vervolgens verder ontwikkeld voor geschreven Nederlandse teksten.

sent. #	word #	word	Tag	lemma
0	0	Ze	VNW(pers,pron,stan,red,3,ev,fem)	ze
0	1	trekt	WW(pv,tgw,met-t)	trekken
0	2	met	VZ(init)	met
0	3	veel	VNW(onbep,grad,stan,prenom,zonder,agr,basis)	veel
0	4	kabaal	N(soort,ev,basis,onz,stan)	kabaal
0	5	een	LID(onbep,stan,agr)	een
0	6	koffer	N(soort,ev,basis,zijd,stan)	koffer
0	7	onder	VZ(init)	onder
0	8	haar	VNW(bez,det,stan,vol,3,ev,prenom,zonder,agr)	haar
0	9	bed	N(soort,ev,basis,onz,stan)	bed
0	10	vandaan	VZ(fin)	vandaan
0	11	en	VG(neven)	en
0	12	opent	WW(pv,tgw,met-t)	openen
0	13	de	LID(bep,stan,rest)	de
0	14	roestige	ADJ(prenom,basis,met-e,stan)	roestig
0	15	sloten	N(soort,mv,basis)	slot
0	16	.	LET()	

Op onze website kunt u online een demo van de MBDDagLem bekijken door losse zinnen in te typen of een korte tekst (maximaal 128 KB) te uploaden. U vindt daar ook meer informatie over de tagset en de memory-based learning technieken. Voor het taggen of lemmatiseren van grotere verzamelingen tekst, kunt u contact opnemen met de TST-centrale via tst@inl.nl.

Colofon

Deze nieuwsbrief is een uitgave van de afdeling TST-centrale van het Instituut voor Nederlandse Lexicologie, Postbus 9515, 2300 RA Leiden.
t +31 (0)715272282 (Leiden) / t +32 (0)38202784 (Antwerpen)
Ontwerp/opmaak: ufork/Swantje Haage Ontwerp, Amsterdam