

# De TST-Centrale en de TST-materialen

Taal in Bedrijf  
November 2008



*De TST-Centrale is een initiatief van de Nederlandse Taalunie, wordt gefinancierd door de Nederlandse Taalunie en is ondergebracht bij het Instituut voor Nederlandse Lexicologie.*



# De TST-Centrale

- De Centrale voor Taal- en Spraaktechnologie is de Nederlands-Vlaamse centrale voor beheer, onderhoud en distributie van Nederlandse digitale taalmaterialen.



# Inhoud

- Missie, verankering
- Takenpakket
- Evaluatie en verbetering
- Met het veld
- Verwacht
- Vragen

# Missie

- Het tegengaan van kapitaalvernietiging door het (her)gebruik van taalmaterialen te stimuleren
  - De (basis)taalmaterialen zijn veelal met overheidsgeld gefinancierd en worden door de TST-Centrale onderhouden en beschikbaar gesteld voor onderwijs, onderzoek en ontwikkeling.

# Verankering

- Initiatief van Nederlandse Taalunie (NTU)
- Gefinancierd door NTU (± 5 fte)
  - Tusschenbesteding: meële organisatie met aandacht voor samenwerking
- Productief Instituut voor Nederlandse Lexicologie
- Operationeel vanaf 2004
- In Leiden en dependance Antwerpen





# Takenpakket

- Acquisitie
- Beheer
- Onderhoud
- Beschikbaarstelling
- Service



# Acquisitie

- Verwerving van taalmaterialen uit o.a. subsidieprogramma's (zoals STEVIN)
- Maken van afspraken met leveranciers over dienstverlening op maat
- Opstellen en afsluiten van licenties met datagebruikers en met dataleveranciers
- Het (helpen) klaren van kwesties inzake intellectuele eigendomsrechten (ipr)



# Acquisitie (2 recente voorbeelden)

- STEVIN
  - Tekstcorpora, spaakcorpora, lexica, tools  
(Autonomata, Corea, D-Coi, IRME, JASMIN)
- Stichting Hebreeuwse en Jiddisje woorden in het Nederlands
  - Sofeer: gratis digitaal woordenboek Hebreeuwse en Jiddisje woorden in het Nederlands  
(voor gebruik in MS Office)



# Beheer

- Opslag & back-up
  - SAN/NAS met  $\pm 1,5$  TB aan taalmaterialen ( $\pm 50$  taalmaterialen)
- Versiebeheer
  - Bijv. Corpus Gesproken Nederlands (CGN): eindrelease, updates, versie 2.0, Corex7, webservice
- Ook (kennis)documentatie
- Volgens best practices: Information Technology Infrastructure Library (ITIL)



# Onderhoud (klein)

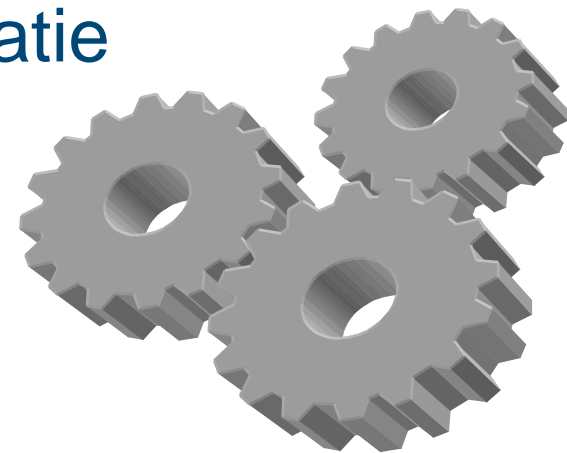
- Structurele taak: bruikbaar houden
- Testen
  - “Werkt X onder Vista/IE7/FF3?”
- Migratie
  - Naar nieuwe(re) hard- en/of software (standaardisatie)
- Bugs inventariseren en (kritieke bugs) verhelpen





# Onderhoud (groot)

- Projectmatig (bij vraag en financiering): wenselijke verbeteringen realiseren
- In samenwerking met het veld
- Bijvoorbeeld
  - Gebruikersinterface en webapplicatie ontwikkelen voor CGN
  - Uitontwikkeling productiestraat Modern Grammar of Dutch





# Beschikbaarstelling

- Informeren
  - Website met productcatalogus en -informatie
    - O.a. vergelijkingstabellen
- Raadplegen
  - Webapplicaties (Mijn TSTC)
- Distributie
  - Online
  - Offline

## Home

Over het INL

Afdelingen >

**TST-Centrale**

Medewerkers >

Bestuur

Begeleidingscommissies >

Studieprogramma

Vacatures en stages



## Nieuwe producten (17-10-2008)

STEVIN-resultaten:

- [AUTONOMATA-q2p-toolkit](#)
- [AUTONOMATA-namencorpus](#)
- [COREA-coreferentiecorpus](#)
- [D-Coi-corpus](#)
- [DuELME](#)
- [JASMIN-spraakcorpus](#)

## Neologisme van de week

Bumperkleven nu nog  
asociaal met de ...

[Lees meer...](#)

## TST-Centrale

De eerste STEVIN-resultaten zijn opgenomen in de [productcatalogus](#). Let op: vanaf 1 november worden handlingkosten ingevoerd voor offline distributie. [Lees meer...](#)



De Centrale voor Taal- en Spraaktechnologie (TST-Centrale) is de Nederlands-Vlaamse centrale voor beheer, onderhoud en distributie van Nederlandse digitale taalmaterialen. De taalmaterialen zijn veelal met overheidsgeld gefinancierd en worden door de TST-Centrale onderhouden en beschikbaar gesteld voor onderwijs, onderzoek en ontwikkeling.

De (informatie over) digitale taalkundige bronnen die de TST-Centrale beheert kunt u vinden in het hoofdmenu onder [Producten](#). Via de Productcatalogus kunt u diverse producten online raadplegen, licenties afsluiten, documentatie bekijken en overige productinformatie vinden. Voor het online raadplegen [registreert](#) u zich eenmalig voor een gebruikersaccount. Voor het bekijken van productinformatie heeft u geen account nodig.

De TST-Centrale is op initiatief van de [Nederlandse Taalunie](#) opgericht in 2004. De TST-Centrale wordt gefinancierd door de Nederlandse Taalunie en is als project ondergebracht bij het INL.

## Inloggen

E-mailadres

Wachtwoord

[Wachtwoord vergeten?](#)

[Registreren](#) | [Waarom?](#)

 TST-Centrale

## Nieuws

- [Invoering verzend- en handlingkosten TST-Centrale](#)
- [STEVIN-resultaten beschikbaar bij de TST-Centrale](#)
- [Halfjaarverslag IMPACT](#)
- [Digitaal woordenboek van Hebreeuwse en Jiddische woorden online](#)
- [INL tekent overeenkomst met Universiteit Leiden](#)

[Meer nieuws...](#)

## Nieuwsbrief

- [INL-nieuwsbrief \(pdf\)](#)
- [TST-nieuwsbrief 3 \(pdf\)](#)
- [Archief / abonneren](#)



# Productinformatie

- Online catalogus
- Thematisch overzicht
- Doorzoekbaar
- Informatie per product
  - Product sheet
  - Prijzen en voorwaarden
  - Documentatie
  - Producent en financier
  - ...

<b>Naam:</b>	AUTONOMATA-g2p-toolkit
<b>Afkorting:</b>	AUTOT
<b>Omschrijving:</b>	De AUTONOMATA-g2p-toolkit bestaat uit een transcriptietool en learning tools, waarmee men woordenlijsten kan verrijken met nauwkeurige uitspraakinformatie. De tool maakt gebruik van een algemene grafeem-naar-foneemomzetter (de g2p-omzetter) en een foneem-naar-foneemomzetter (de p2p-omzetter) die voor specifieke domeinen (zoals persoonsnamen en geografische namen) de resultaten van de g2p-omzetter verbetert.
<b>Bestellen:</b>	neem contact op met de <a href="#">servicedesk</a>
<b>Servicedesk:</b>	<a href="#">reactieformulier</a>
<b>Prijzen (excl. btw):</b>	€ 500 (niet-commercieel gebruik, geheel) op aanvraag (commercieel gebruik, geheel)
<b>Product sheet:</b>	<a href="#">pdf</a>
<b>Documentatie:</b>	<a href="#">overzichtsartikel</a> : ontwikkeling uitspraaklexicon voor namen (2008) <a href="#">Interspeech2007-artikel</a> : g2p-conversie van namen <a href="#">LREC2006-artikel</a> : ontwikkeling p2p-converter voor verbetering g2p-conversie van namen
<b>Versie/jaar:</b>	2.0, 2008
<b>Distributievorm:</b>	download
<b>Producent:</b>	ELIS - UGent, CLST - RU Nijmegen, UiL - OTS Utrecht, Nuance, TeleAtlas
<b>Financier:</b>	NTU STEVIN
<b>Gerelateerd:</b>	<a href="#">AUTONOMATA-namencorpus</a>



# Vergelijkingstabellen

- Voor corpora, monolinguale en bilinguale lexica
- Zie poster

		gesproken/ geschreven bronnen	gesproken bronnen		
		Eindhoven Corpus	Corpus Gesproken Nederlands	IFA corpus	IFA Dialog Video Corpus
corpusinhoud	periode	1960-1973	1991-2003	2001	2006
	aantal woorden	720.000	9.000.000	50.000	[5 uur spraak]
	geluidsbestanden	-	+	+	+
	video-opnames	-	-	-	+
spraak	spontaan (face-to-face, interviews, telefoongesprekken etc.)	+	+	+	+
	voorbereid (voorgelezen verhalen/ zinnen, (nieuws-) uitzendingen etc.)	-	+	+	-

	Monolinguale computationele lexica					
	Bronbestand Woordenlijst Nederlandse Taal 2005	e-Lex		Referentie- bestand Nederlands	Referentie- bestand Belgisch- Nederlands	PAROLE- lexicon
		Enkelwrđ- lexicon	Meerwrđ- lexicon			
lemma's	110.000	220.000	26.000	45.000	4.000	20.000
woordvormen		620.000	77.000			
orthografie	++	+	+	+	+	+

	Bilinguale computationele lexica		
	OMBI-Dutch-Arabic	OMBI-Dutch-Indonesian	OMBI-Dutch-Danish
talen	Nederlands / Arabisch	Nederlands / Indonesisch	Nederlands / Deens
lemma's	36.000	46.000	45.000
orthografie	+	+	+



# Webapplicaties

- Ieder taalmateriaal online raadpleegbaar
- Basis: *rapid web application development framework* BOB voor een eerste indruk
- Geavanceerd: productspecifieke webapplicaties
  - Indien beschikbaar of uit groot onderhoud
- Benaderbaar via Mijn TSTC



# Mijn TSTC

- *Winkelmandje* voor toegang tot uw favoriete online taalmaterialen
- Eenmalige registratie
- Zoveel mogelijk uniforme *look and feel*
- Nieuws en informatie via onze nieuwsbrief
- Wij: goed zicht op gebruik en gebruikers

**Inloggen**

E-mailadres

Wachtwoord

[Wachtwoord vergeten?](#)  
[Registreren](#) | [Waarom?](#)

**rū | TST-Centrale**

**Inloggen**

U bent ingelogd als  
**veenendaal@inl.nl**

Via de [productcatalogus](#) kunt u gebruikmaken van de online producten waarvoor u een licentie heeft aangevraagd.

[Gegevens en licenties](#)

[Log uit](#)

**rū | TST-Centrale**



# Webapplicatie (BOB)

- Referentiebestand Belgisch Nederlands (RBBN)

Uitloggen

Zoek Opnieuw

◀ bladeren ▶

1 van 1

Referentiebestand Belgisch-Nederlands

[item](#) aanwerven

[sorteerlemma](#) aanwerven

[volgnummer](#)

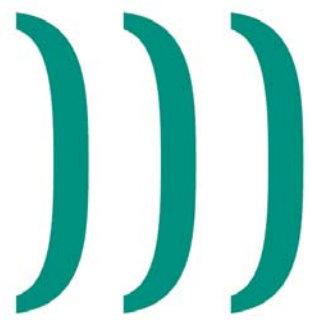
voor een toelichting op de velden en hun mogelijke waarden, [klik hier](#)

## lexicale informatie

<a href="#">woordklasse</a>	WW
<a href="#">type item</a>	w
id betekenisid	1005
<a href="#">soort belgicisme</a>	vrije alternant
<a href="#">definitie</a>	in dienst nemen (personeel)
<a href="#">bron definitie</a>	gvd
<a href="#">selectie BW</a>	1

## methodologische informatie

<a href="#">NRC freq</a>	0	<a href="#">cultuurgebonden</a>		<a href="#">variant</a>	in dienst nemen
<a href="#">roul freq</a>	40	<a href="#">domein</a>			
<a href="#">Google.nl</a>		restricties			
<a href="#">Google.be</a>		<a href="#">type restrictie</a>		<a href="#">lexicalisatie van variant</a>	ja
<a href="#">CGN- freq</a>		<a href="#">waarde restrictie</a>		<a href="#">syntaxis</a>	0
<a href="#">gvd label</a>	b	<a href="#">informanten bron</a>		<a href="#">roul freq van variant</a>	40
<a href="#">restr qvd</a>		<a href="#">a score trefwoord</a>		<a href="#">a score variant</a>	
<a href="#">verschuieren</a>		<a href="#">b score trefwoord</a>		<a href="#">b score variant</a>	
<a href="#">verschuieren label</a>					



# Webapplicatie (geavanceerd)

- Woordenboek der Nederlandsche Taal online (iWNT)

## CENTRALE

Woordsoort: znw.(m.,v.)  
Modern lemma: centrale

Koppelingen:  
Vorig artikel: CENTRAALKRACHT  
Volgend artikel: CENTRALE-  
VERWARMINGSBUIS

[sluit](#)

znw. m. en vr., mv. *-s*, soms *-n*. Substantivering van *centraal*. Fr. *central(e)* (1883, m., in de bet. 'telefooncentrale'; 1926, vr., in de bet. 1; 1956, vr., in de bet. 2), amer.-eng. *central* (1889 in de bet. 'telefooncentrale'), *du. zentrale* (2de h. 19de e. in de bet. 2).

- 1.** Electriciteitsfabriek als centrum van het electriciteitsnet.  
2°. Installatie op een schip ter opwekking en verspreiding van electriciteit.  
ODERWALD, *Wdb. Scheepsdienst* [1931].  
— Hiermede is de inrichting der machine- en ketelruimen (*van een schip*) nog niet ten einde. Een onderdeel, dat de laatste jaren meer en meer toepassing vindt, is de elektrische installatie, die van een kleine huisinstallatie tot een behoorlijke elektrische centrale is uitgegroeid, WILKE e.a., *Scheepsb.* 144 [1944].
- 2.** Instelling of installatie als centraal punt van waaruit goederen, diensten, berichten e.d. verspreid worden. Dikwijls als verkorting, b.v. uit *telefooncentrale*.



# Online distributie

- Via Mijn TSTC toegang tot downloadbare taalmaterialen
  - Qua grootte geschikt voor downloaden
- Incidenteel distributie via e-mail
  - Enkele kleine materialen
- Plannen: volwaardige *webwinkel*

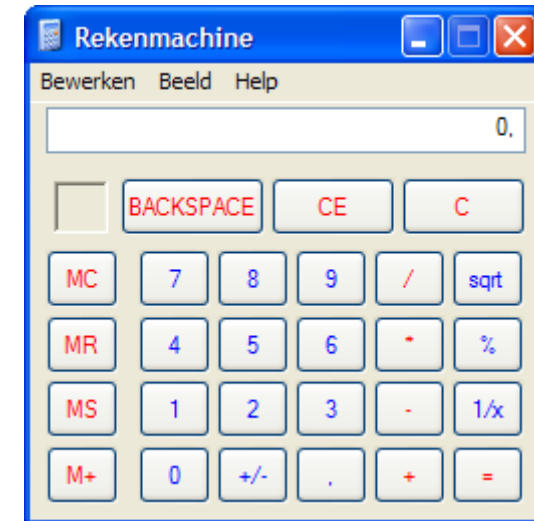


# Offline distributie

- Afhankelijk van grootte taalmateriaal
  - Op dvd
  - Op harde schijf
- Beleid (i.s.m. Prijzencommissie):
  - Gratis bij niet-commercieel gebruik
  - Marktconform bij commercieel gebruik
  - Vergoeding voor verzenden en handling

# Enkele cijfers...

- Licenties voor offline distributie
  - 2006 23 (o.a. 8 onderzoek, 4 commercieel)
  - 2007 31 (o.a. 19 onderzoek, 4 commercieel)
  - 2008... 34 (o.a. 28 onderzoek, 3 commercieel)
- Licenties voor online diensten
  - 2006  $\pm 600$
  - 2007  $\pm 50.000$  (onlinegang iWNT)
  - 2008...  $\pm 60.000$
- Meldingen verwerkt door servicedesk
  - 2006  $\pm 100$  (maar ook decentraal)
  - 2007 1400 (onlinegang iWNT)
  - 2008... 1300





# Dienstverlening

- Ondersteunende en adviserende servicedesk in samenwerking met experts uit het veld
- Invoering *ITIL-compliant* servicedeskpakket in 2009
- Webcursussen, workshops, gastcolleges
- Kennismanagement



# Evaluatie en verbetering

- Diepte-, breedte- en zelfevaluatie
  - In opdracht van NTU, in 2007
  - Bestaan TST-Centrale waardevol
  - Gebruikers tevreden, maar aandachtspunten vanuit partners en leveranciers
  - Actieplan in ontwikkeling
- Opnieuw evaluatie in 2009

# Met het veld (1/2)

- Contact met het veld
  - Commissies (STEVIN, ALVV, CoTerm, ...)
  - Deelnemer en bestuurslid NOTaS
  - Deelname aan events
  - “De TST-Centrale komt naar je toe”
- Samenwerking met het veld
  - Acquisitie, klein en groot onderhoud, servicedesk
  - Ook: communicatiematerialen en juridisch advies

# Met het veld (2/2)

- Sponsoring
  - LREC, CLIN, TLT ('08), TABU ('09)
- Projecten
  - Distributed Access Management for Language Resources (DAM-LR, [www.dam-lr.eu](http://www.dam-lr.eu))
  - Common Language Resources and Technology Infrastructure Network (CLARIN, [www.clarin.eu](http://www.clarin.eu))



26/29



# Verwacht...

Updates van het bestand neologismen online; ANW-corpus online (100 miljoen woorden tekstcorpus); Terminologische lexicons en software (Termextractor); onlineversie CGN; Regionale (dialect)woordenboeken; STEVIN Daeso (corpus en software voor semantiek); STEVIN DPC (parallele corpora NI-En en NI-Fr); STEVIN Lassy (syntactisch geannoteerd corpus); STEVIN Midas (software voor robuuste spraakherkenning); STEVIN N-best (benchmark voor Nederlandstalige spraakherkenning); STEVIN-can-Praat (software voor spraakonderzoek); STEVIN Spraak (spraakherkenner); STEVIN Cornetto (softwaretoolkit voor lexicale semantiek); diverse historische tekstcorpora; updates en uitbreidingen van het WNT; update van het Referentiebestand Nederlands; (updates van de) CLVV/ALVV-bilinguale woordenboeken; Modern Grammar of Dutch; ...



# Kortom

- Voor u als gebruiker
  - Hét loket voor digitale taalmaterialen en dienstverlening
- Voor u als leverancier
  - Agentschap voor uw taalmaterialen
    - Afspraken over beheer, onderhoud, distributie en dienstverlening
- Voor u als partner
  - Mogelijkheden tot samenwerking bij bijvoorbeeld onderhoud en dienstverlening



# Vragen?

- Plenair: nu



- Persoonlijk/producten/overige: bij posters
  - Anna, Griet, Laura, Michel, Remco