TST-Centrale

Zoekacties en gebruikte codes

binnen

COREX (CORpusEXploratiesoftware)

➔ Vragen, opmerkingen of suggesties: servicedesk@inl.nl

➔ Informatie over het CGN: http://www.inl.nl/tst-centrale/nl/producten/corpora/corpus-gesproken-nederlands/6-17

➔ CGN-webcursus: http://www.inl.nl/tst-centrale/images/stories/producten/webcursus/cgn_webcursus/

1 Metadata search

1.1 Session Keys (via Session)

CGN.wordCount	number of words in the sample {number}
CGN.recCount	duration of the sample expressed in the total number of seconds {number}
CGN.recDate	recording date {(date or) year}
CGN.fon.available	whether there is a manually verified phonetic annotation available or not {true, false}
CGN.syn.available	whether there is a syntactic annotation available or not {true, false}
CGN.pro.available	whether there is a prosodic annotation available or not {true, false}

1.2 Participant Keys (via Session/Participant)

CGN.age	age class to which speaker belonged at the time the sample was recorded;	
	age0	under 18 years of age
	age1	18-24 years
	age2	25-34 years
	age3	35-44 years
	age4	45-55 years
	age5	over 55 years of age
	ageX	age unknown
CGN.birth.year	speaker's year of birth	
CGN.birth.place	speaker's place of birth, represented in terms of the first three digits of the postal code preceded by the country code (BE, NL)	
CGN.birth.region	(geographical) region where the speaker was born. For a list of regions distinguished, see below.	
CGN.residence.place	speaker's place of residence, represented in terms of the first three digits of the postal code preceded by the country code	
CGN.residence.reg	(geographical) region where the speaker resides. For a list of regions distinguished, see below.	
CGN.residence.size	indication of the (present) size of the place where the speaker resided while (s)he was between 4 and 16 years of age	
CGN.education.place	place where speaker lived for the most part between ages 4 and 16 represented in terms of the first three digits of the postal code preceded by the country code (BE, NL)	
CGN.education.opleiding	type of education; e.g. lager onderwijs, mbo, universiteit	
CGN.education.reg	(geographical) region where speaker lived while (s)he attended secondary education (region where speaker lived for the most part between ages 4 and 16). For a list of regions distinguished,	

	see below.	
CGN.education.level	level of education	
	edu1	high
	edu2	middle
	edu3	low
	edu4	unknown

1.3 Geographical regions:

regN1a	NL, central region, Zuid-Holland, excl. Goeree Overflake
regN1b	NL, central region, Noord-Holland, excl. West-Friesland
regN1c	NL, central region, West-Utrecht, incl. the city of Utrecht
regN2a	NL, transitional region, Zeeland, incl. Goeree Overflakee and Zeeuws-Vlaanderen
regN2b	NL, transitional region, Oost-Utrecht, excl. stad Utrecht
regN2c	NL, transitional region, Gelders rivierengebied, incl. Arnhem and Nijmegen
regN2d	NL, transitional region, Veluwe up to the river IJssel
regN2e	NL, transitional region, West-Friesland
regN2f	NL, transitional region, Polders
regN3a	NL, peripheral region 1 (north east), “Achterhoek”
regN3b	NL, peripheral region 1 (north east), Overijssel
regN3c	NL, peripheral region 1 (north east), Drenthe
regN3d	NL, peripheral region 1 (north east), Groningen
regN3e	NL, peripheral region 1 (north east), Friesland
regN4a	NL, peripheral region 2 (south), Noord-Brabant
regN4b	NL, peripheral region 2 (south), Limburg
regNx	NL, unknown
regV1	Flanders, central region (Antwerpen and Vlaams-Brabant)
regV2	Flanders, transitional region (Oost-Vlaanderen)
regV3	Flanders, peripheral region 1 (West-Vlaanderen)
regV4	Flanders, peripheral region 2 (Limburg)
regVx	Flanders, unknown
regW	Wallonia
regZ	region known to be outside of The Netherlands and Flanders
regX	region unknown

1.4 Content Keys (via Session/Content/Keys)

CGN.textclass.target	gives information about four aspects: text type (specifies the component to which a sample belongs), degree of preparedness, mode and domain	
	Text types	
	tta	spontaneous conversations (face-to-face)
	ttb	interviews with teachers of Dutch
	ttd	spontaneous telephone dialogues (recorded on MD with local interface)
	tte	simulated business negotiations
	ttf	interviews/discussions/debates (broadcast)
	ttg	(political) discussions/debates/meetings (non-broadcast)
	tth	lessons recorded in a classroom
	tti	live (e.g. sport) commentaries (broadcast)
	ttj	news reports/reportages (broadcast)
	ttk	news (broadcast)
	ttl	commentaries/columns/reviews (broadcast)
	ttm	ceremonious speeches/sermons
	ttn	lectures/seminars
	tto	read speech
	Degree of preparedness	
	prep1	scripted,
	prep2	unscripted,
	prep3	more-or-less scripted;
	mode	
	mod1	broadcast, radio;
	mod2	broadcast, tv;
	mod3	non-broadcast;
	Domain	
	dom1	private
	dom2	public

2 Content Search

2.0 Reguliere expressies

Het is binnen de Corex-tool “content search” mogelijk om je zoekactie te definiëren met behulp van een reguliere expressie. Een **reguliere expressie** is een manier om patronen te beschrijven waarmee een computer tekst kan herkennen. Dat is handig wanneer je bijvoorbeeld wilt zoeken op alle woorden die met “over“ beginnen, zonder dat je al die woorden afzonderlijk moet invoeren.

Variabele	**Betekenis**
.	matches an arbitrary character
+	matches the preceding pattern element one or more times
?	matches the previous character zero or one times
*	matches the previous character zero or more times
^	matches the beginning of a word
$	matches the end of a word
[adg]	matches the character a, d or g
[^ad]	matches any character except a and d

BELANGRIJK: Wanneer je een zoekstring ingeeft, dan zoekt Corex niet exact op die zoekstring, maar naar alle woorden die die zoekstring bevatten. Het is niet aan te raden om variabelen zoals “.*” voor of na je zoekstring te zetten, want dat levert verkeerde resultaten op.
Wanneer je wil aangeven dat je zoekstring aan het begin of aan het einde van een woord moet voorkomen, kun je gebruikmaken van de tekens ^ en $.

Voorbeelden van zoekacties wanneer “Reg. Exp.” aangevinkt is:

- **bel** vindt bel, gebeld, bellen, babbelen, dubbel,…
- **^bel** vindt bel, bellen, belt, belangrijke,…
- **bel$** vindt bel, dubbel, heibel,…
- **^b.l$** vindt bel, bil, bol, bal,…
- **^b.*l$** vindt bel, bil, bol, bal, bedoel, bewijsmateriaal, boksschool,…
- **^b[ae]l$** vindt bel en bal
- **^.it$** vindt uit, zit, dit,…

2.1 Orthography, marked, lemma

Codering	**Betekenis**
Dialect	Dialectwoord
Foreign	Vreemd/buitenlands woord
Incomplete	Afgebroken woord
Mispr	Verkeerd uitgesproken woord of klanknabootsing
Regionalpr	Zwaar dialectisch uitgesproken woord
uncertain	Woorden waarvan de transcribent niet zeker is of hij/zij ze goed gehoord heeft

2.2 Fonetische transcriptie

2.2.1 CGN-symbolen

Klasse	Voorbeeld	CGN-symbool
Plosieven	put	p
	bad	b
	tak	t
	dak	d
	kat	k
	goal	g
Fricatieven	fiets	f
	vat	v
	sap	s
	zat	z
	sjaal	S
	ravage	Z
	licht	x
	regen	G
	geheel	h
Sonoranten	lang	N
	mat	m
	nat	n
	oranje	J
	lat	l
	rat	r
	wat	w
	jas	j

Klasse	Voorbeeld	CGN-symbool
Korte vocalen	lip	I
	leg	E
	lat	A
	bom	O
	put	Y
Lange vocalen	liep	i
	buur	y
	leeg	e
	deuk	2
	laat	a
	boom	o
	boek	u
Sjwa	gelijk	@
Diftongen	wijs	E+
	huis	Y+
	koud	A+
Leenvocalen	scène	E:
	freule	Y:
	zone	O:
Nasale vocalen	vaccin	E~
	croissant	A~
	congé	O~
	parfum	Y~

2.2.2 Zoekcodes voor de fonetische transcripties

Code	Betekenis
SEP	gescheiden (er wordt geen foneem gedeeld over de woordgrenzen heen)
SHARE-P	een gedeeld plosief (plofklank) over woordgrenzen heen
SHARE-NP	een gedeeld non-plosief over woordgrenzen heen
SHARE-W	uitzondering: het woord/de woorden ‘da’s’
INSERT	een geïnserteerd foneem tussen twee woorden

2.3 Part of speech

Afkorting woordsoort	Woordsoort
ADJ	Adjectief
BW	Bijwoord
LET	Leesteken
LID	Lidwoord
N	Substantief
SPEC	speciale woorden (vb. delen van eigennamen, woorden in zelfnoemfunctie,…)
TSW	Tussenwerpsel
TW	Telwoord
VG	Voegwoord
VNW	Voornaamwoord
VZ	Voorzetsel
WW	Werkwoord

Het is mogelijk om binnen de content search te zoeken op woordsoort (vb. WW) of op een volledige tag (vb. N(soort,mv,basis), maar het is ook mogelijk om een zoekactie te doen op basis van gedeeltelijke POS-tags.

Voorbeeld: er zijn meerdere tags voor onbepaalde lidwoorden, nl. LID(onbep, dial), LID(onbep,gen,evf) en LID(onbep,stan,agr).
Om alle uitingen te vinden waarin onbepaalde lidwoorden voorkomen, kun je drie afzonderlijke zoekacties doen, maar het is ook mogelijk om met één zoekopdracht alle uitingen met onbepaalde lidwoorden te vinden.

Dit kan door met de rechtermuisknop te klikken op het vak na LID waarin je reeds een van de tags voor onbepaalde lidwoorden geselecteerd hebt, bijvoorbeeld (onbep,stan,agr). Na de klik wordt het vak wit en kan je de cursor in het vakje plaatsen. Plaats de cursor vóór dat deel van de tag dat je wil deleten, bv. voor “stan”. Vervolgens druk je op de deleteknop en dan wordt de tag vanaf “stan” vervangen door een sterretje: *. Er staat nu als tag LID(onbep,*). Wanneer je een zoekactie start worden alle onbepaalde lidwoorden gezocht.

Let op: alleen het laatste deel van een tag kan verwijderd worden.

2.4 Prosodische annotatie

De volgende prosodische fenomenen maken deel uit van de prosodische annotatie:

- **prosodische grenzen** (een onderbreking in het spraaksignaal):
    - zwakke grenzen zijn per definitie aangeduid in een manuele annotatie en binnen Corex kregen die de naam **weak**
    - sterke grenzen werden aangeduid in een manuele annotatie (strong) of in een automatische annotatie (strong)
- **klemtoon**
- **abnormale klankverlenging**

De volgende zoekacties zijn mogelijk binnen de prosodische annotatie (zie ook hieronder):

- je kan bepalen of de grens die je zoekt links of rechts van een woord moet staan (leftb en rightb)
- je kan op een bepaald aantal grenzen zoeken (nweakb en nstrongb)
- je kan zoeken op een bepaald aantal klemtonen (nprom) al dan niet in combinatie met een woord
- je kan zoeken op een bepaald aantal klankverleningen (nlength) al dan niet in combinatie met een woord