

Zoeken in het GCND-corpus met XPath

Zie ook * <https://hackmd.io/@amghysel/r1kMS8cC9> * Voor een algemeen GrE-TEL tutorial zie <https://surfdrive.surf.nl/files/index.php/s/xfjVB2AfwgOpmNM>

* Ook <https://paqu.let.rug.nl:8068/info.html#re>

Motivatie waarom het in veel geval noodzakelijk is met XPath aan de slag te gaan:

- De example-based search van GrETEL zal voor sommige dialectconstructies niet goed werken omdat Alpino de gebruikersinvoer niet op de gewenste manier analyseert. Zie bijvoorbeeld subjectverdubbeling (1.1)
- In de example-based search kom je niet meteen tot de essentie van wat je zoekt; om een hogere *recall* te bereiken zal de gegenereerde query moeten worden aangepast. Hieronder een voorbeeld:

We zoeken naar “groter dan/of/als X”-constructies. We voeren “*groter dan een olifant*” in bij de example-based search.

De analyse is:

en de bijbehorende xpath is

```
//node[@cat="ap" and @rel="--" and
  node[@pt="adj" and @rel="hd"] and
  node[@cat="cp" and @rel="obcomp" and
    node[@pt="vg" and @rel="cmp"] and
    node[@cat="np" and @rel="body" and
      node[@pt="lid" and @rel="det"] and
      node[@pt="n" and @rel="hd"]]]]
```

Hiermee worden 2 resultaten gevonden - een teleurstellend resultaat. Deze query is duidelijk te restrictief. Naar `@rel="--` waren we niet op zoek, en eigenlijk maakt de vorm van het vergelijkende element ook niet uit. We moeten de query dus tot zijn essentie reduceren:

```
//node[@cat="ap" and
  node[@pt="adj" and @rel="hd"] and
  node[@cat="cp" and @rel="obcomp" and
    node[@pt="vg" and @rel="cmp"] and
    node[@rel="body" ]]]
```

Hiermee vinden we 118 resultaten, een aannemelijker aantal.

1. Subjectsverschijnselen

1.1 subjectverdubbeling (of drievoudig subject)

- *Ik heb ik ik dat niet gezegd.*
- *en t jij ee t jij zijn kazak gekeerd .*

```
//node[count(./node[@rel='su']) > 1]
```

1.1 subject in objectvorm

omdat hem peinsde dat dat zijn kindje was.

```
//node[@rel="su" and @word="hem"]
```

Iets algemener:

```
//node[@rel="su" and @naamval="obl"]
```

1.3 Presentatief ‘het’

Deze constructie wordt noch in de Lassy-handleiding noch in de CGN-beschrijving besproken. In het GCND is dit *het* zoals presentatief *er* behandeld, en heeft *het* dus het dependentielabel *MOD*.

?Adjectieven met “geen”.

Vindbaar met:

```
//node[@rel='mod' and word='het']
```

De resultaten zijn soms een beetje verwarrend

2. Uitbreidingen van de zin: TAG en SAT

Discourse-structuren in de Lassy annotatie

Uit het Lassy-annotatiemanual:

	OMSCHRIJVING
DEPEND	discourse-unit
NUCL	kernzin
DLINK	discourse-link In samenhangende tekst komt het vaak voor dat een spreker een zin begint of eindigt met iets dat op een voegwoord lijkt, maar geen syntactisch verband heeft: <i>maar wat ik zeggen wou . . .</i>
SAT	satelliet: aan- of uitloop ‘aanloop’- en ‘uitloop’-elementen die een anaforische relatie onderhouden met verwijswoorden in de kernzin. <i>NUCL:[het verbaast me] SAT:[dat je dat nog weet]</i>
TAG	aanhangsel, tussenvoegsel. Elementen die op een andere manier dan DLINK en SAT buiten de kernzin staan
DP	discourse-part

categorielabel	OMSCHRIJVING
DU	discourse-unit

2.1 Linker zinsperiferie

2.1.a Aanloopconstructie (Left dislocation) *Jan, die ken ik niet*

Dit goed te gaan in Alpino, en kan dus via example-based search worden gevonden.

Herkenbaar aan dependentierelatie *SAT* en (categorie *np* of woordsoort zelfstandig naamwoord (*n*)).

```
//node[@rel='sat' and (@cat='np' or @pt='n')][@begin="0"]
```

Niet altijd makkelijk te onderscheiden van volgende categorie.

2.1.b Hanging Topic / Hangend Topic / Nominativus Pendens: *mijn vent wist hij ook niet wat dat was en nu komt ...*

Er staat steeds een naamwoordgroep in de eerste positie, die later in de zin door een persoonlijk voornaamwoord (hij, het, zij, hem, haar) wordt opgenomen

Nominale tag-nodes aan het begin van de zin zoek je met

```
//node[@rel='tag' and (@cat='np' or @pos='noun') and @begin="0"]
```

Niet alle matches van deze query zijn daadwerkelijk topicalisaties.

2.1.c Tussenwerpsels en aansporingen

- *zo, dat was plezant.*
- *natuurlijk, moeilijk is het niet.*
- *kom, ik ga er maar vandoor.*
- *jongens, ik vertrek nu.*

Met dit soort structuren kan Alpino doorgaans vlotjes om; preprocessing is dan ook niet nodig.

Geanalyseerd met dependentierelaties tag (voor tussenwerpsel of aansporing) en nucl (voor de eigenlijke zin)

```
//node[@rel='tag' and (@cat="pp" or @pt='bw' or @cat="advp" or @pt="tsw") and @begin="0"][.].
```

2.1.d Inversieloos V-later-dan-2 / V>2 / Noninverted V3 *zeg als je nu trouwt het zijn altijd voort kosten.*

```
//node[@rel="tag" and @cat="cp"]
```

Mogelijk ook:

```
//node[@rel="tag" and @cat="pp"]
```

in de zomer t e klaar tot sn avonds t negenen

2.1.e Ingebedde dislocaties Wat vindt u der eigenlijk van dat zulke zinnen dat die zo geanalyseerd worden?

Zijn getagd met met SAT

2.1.f ja/nee het/ik/...

- *Bwa nee het jong*
- *ja **ja** ze het is heel juist*

```
//node[@rel='tag'] [node[@rel='mwp' and @pt='tsw'] and node[@rel='mwp' and @pos='pron']]
```

2.2 V2-bijzinnen - pseudodirecte rede

Het is toch geen waar, etc

- *hij zei hij weet het niet*
- *ik zeg gisteravond , ik moet de auto binnensteken ut tut tut .*
- *ik zeg , steek hem binnen .*

(Laatste met sv1, verschil met smain niet zo duidelijk?)

- Inleidende matrixzin (hij zei):

Dependentielabel (rel): tag Categorielabel (cat): smain

Pseudodirecte rede - V2-bijzin (hij weet het niet):

Dependentielabel (rel): nucl

Categorielabel (cat): smain (of - bij werkwoordsinitiële zinnen - sv1)

```
//node[./node[@rel='tag' and @cat='smain']] and node[@rel='nucl' and (@cat='smain' or @cat='sv1')]
```

NB: Alpino parset directe en pseudodirecte redes doorgaans automatisch juist als je een komma toevoegt tussen de matrixzin en de V2-bijzin.

2.3 Intercalaties/parentheses/interpositio

Let op: afwijking van Lassy: In het GCND kiezen we ervoor parentheses het dependentielabel TAG te geven en op hetzelfde niveau als de hoofdzin onder te brengen .

```
//node[@rel='tag' and @cat='smain']  
  [number(../node[@cat='smain' and @rel='nucl' and @begin and @end]  
/ @begin) < @begin]  
  [number(../node[@cat='smain' and @rel='nucl' and @begin and @end]/ @end) > @begin]
```

3. Complementizer-fenomenen

3.1 Afwijkende comparatieve voegwoorden (of, als, gelijk als, gelijk of dat)

- *maar het scheelt meer **of de helft** .*

- *dat is veel langer **als dat** ik ik ben .*

Voor ‘of’ bijvoorbeeld:

```
//node[@rel='obcomp'] [./node[@rel='cmp' and @word='of']]
```

Meerwoordige voegwoordelijke combinaties:

```
//node[@rel='obcomp'] [./node[@rel='cmp' and @cat='mwu']]
```

3.2 Directe rede ingeleid door van

- *ja die zeggen van , als we daar in de tranchée en zaten ...*

Vindbaar met:

```
//node[@rel="vc" and @cat="svan"]
```

Bijvoorbeeld beperkt tot combinatie met “zeggen”

```
//node[node[@rel="hd" and @lemma="zeggen"] and node[@rel="vc" and @cat="svan"]]
```

3.3 Expletief dat

Type 1: na onderschikkend voegwoord

- Ik weet niet of dat hij komt.
- Om het te zeggen gelijk of dat het is: ...
- ik was getrouwd sinds dat hij nog bij het leger was

```
//node[@cat='cp']/node[@rel='cmp' and @cat='mwu'] [./node[@word="dat"]]
```

Type 2: na vraagwoord

- Ik weet niet wie dat er komt.
- we gaan weer moeten de tijd afwachten wat dat er allemaal gaat voorvallen

```
//node[@word="wie" and @rel="whd"] [following-sibling::node[./node[@word="dat" and @pt="vg"]]]
```

Type 3: na betrekkelijk voornaamwoord

- *De mens die dat jou moet helpen, zal vloeken.*
- *nee ze voor de oorlog veertien achttien was waren er dan nog knechten **die dat** we winter zomer hadden*

```
//node[@word="die" and @rel="rhd"] [following-sibling::node[./node[@word="dat" and @pt="vg"]]]
```

Type 4: na vraagwoord + of (zeldzaam in Vlaanderen, cf. Lassy-handleiding)

- *Zijn er meer mogelijkheden dan wat of dat je nu hebt?*

(Niet te vinden in corpus)

```
//node[@cat="whsub" and @rel="body" and
  node[@lemma="wat" and @pt="vnw" and @rel="whd"] and
  node[@cat="cp" and @rel="body" and
    node[@cat="mwu" and @rel="cmp" and
      node[@lemma="of" and @pt="vg" and @rel="mwp"] and
      node[@lemma="dat" and @pt="vg" and @rel="mwp"]]]]]
```

3.4 Beknpte bijzinnen ingeleid door *voor* of *van* in plaats van *om*

- *een restaurant voor te blijven voor te eten*

```
//node[@cat='oti'] [./node[@rel='cmp' and @pt='vz' and (@word='voor' or @word='van')]]
```

3.5. Afhankelijke ja/nee-vragen ingeleid door *als* ipv of

- *k weet nie a je da weet .*

```
//node [
  node[@rel="vc"]
  [node[@lemma="als" and
    node[@rel="body"]]] and
  node[@rel="hd" and @pt="ww"]
]
```

Trage query, 23 resultaten voor nu, allemaal west vlaanderen

3.6. Bijzin met hoofdzinsvolgorde (V2-bijzin of Nebensätze)

- *Die rol heb ik heel graag gespeeld omdat er zat poëzie in.*
- *awaar , da zij smokkelden patatten en ...*

Hoofdzinsvolgorde wordt gekenmerkt door * object na werkwoordelijk hoofd * of subject na werkwoordelijk hoofd * Let op object kan in VC zitten * Let op *omdat zij wil broodjes eten* etc zijn weliswaar te duiden als hoofdzinsvolgorden, maar ook als WW-clusteronderbrekingen

Object is losstaand znw (dus geen VC node aanwezig in boom):

```
//node[@cat='ssub'] [
  node[@rel='hd' and @pt='ww'] [number(../node[@rel='obj1' and @word and @pt='n']/@begin) > n
]
```

Object zit binnen VC (dit overlapt met de vlaamse clusterdoorbreking)

```
//node[@cat='ssub'] [
  node[@rel='hd' and @pt='ww'] [number(../node[@rel='vc'] [node[@rel="obj1" and @pt="n"]]/@begin
]
```

Subject na werkwoordelijk hoofd:

```
//node[@cat='ssub']
  [node[@rel='hd' and @pt='ww'] [number(../node[@rel='su'] [1]/@begin) > number(@begin)]]
```

Lastig even later meer doorklooien

```
declare default element namespace "http://alpino.fake.url";
for $node in //node[@cat='ssub'][not (./node[@index])]
    [node[@rel='hd' and @pt='ww'][count(..//node[@rel='su']) = 1][number(..//node[@rel='su']) = 1]]
let $sentence := $node/ancestor::*[local-name()='alpino_ds']/sentence,
    $txt := string-join($node//@word, ' ')
return <node>{$node} <text>{$txt}</text> {$sentence}</node>
```

4. Negatieverschijnselen (o.a. negatiepartikel en en dubbele negatie)

4.1. Dubbele negatie

- (a) Ik en heb dat niet gezegd.
- (b) Ik heb niemand niet gezien.
- (c) Ik heb niets niet gedaan.
- (d) Ik heb dat nooit niet gedaan.
- (e) Daar zijn nooit geen rozen.
- (f) Ik heb geen boeken niet meer.
- (g) Er zijn er niet veel niet meer.
- (h) Ik heb niet veel geen boeken meer.

Negatie met het oude negatiepartikel en (zin a) Alpino ziet *en* standaard als voegwoord.

Negatie met *en* is terug te vinden met een xpath als

```
//node[./node[@rel='mod' and @word='en' and @pt='bw']]
    • ze en hebben geen redenen van klagen
//node
    [./node[@rel='mod' and @word='en' and @pt='bw']]
    [node[@cat='np'][node[@rel='det' and @lemma='geen' and @pt='vnw']]]
```

Negatieverdubbeling binnen de nominale constituent (zin h) Is behandeld als een meerwoordige determiner.

Complexe determiners waar *niet* deel van is, zijn te zoeken met

```
//node[@rel="det" and @cat="mwu"]
    [node[@lemma="niet"]]
```

4.2 Adjectieven die met ‘geen’ gecombineerd worden

- Dat is *geen waar*

!NB zou moeten zijn

```
node[node[@rel='hd' and @pt='ADJ'] and node[@rel='det' and lemma='geen']]
```

Maar in de het corpus heeft in ieder geval *waar* vaak de n-tag,

4.3 Doen-replieken

A: *Hij komt toch niet?* B: *Ja hij en doet ne komt.*

Positieve positieve en negatieve replieken zijn vindbaar met iets als

```
//node[@lemma="doen" and @pvtijd='tgw']  
  [parent::node[@cat='smain']]  
  [../node[@rel="su" and @pt="vnw"]]  
  [not (../node[@rel="obj1"])]  
  [not (../node[@rel="vc" or @rel="predc"])]
```

Negatieve gevallen met *en*

- *bè ik en doe , zei dat kind*

```
//node[@lemma="doen" and @pvtijd='tgw']  
  [../node[@word='en' and @rel="mod" and @pt="bw"]]  
  [parent::node[@cat='smain']]  
  [../node[@rel="su" and @pt="vnw"]]  
  [not (../node[@rel="obj1"])]  
  [not (../node[@rel="vc" or @rel="predc"])]
```

5 Andere

5.1 en zo/ of zo / en al / en alles / maar ja / en si en la

Behandeld als een multi-word unit (MWU) die als modificeerder fungeert (MOD).

6.2 woordherhaling

Het woord wordt 2x opgenomen, met hetzelfde dependentielabel.

6.3 Spiegelconstructies

- je gebruikt nog alsan diezelfde potees gebruik je.
- het is verder is het.
- Ik zeg :“je bent gek”, zeg ik.

Hier worden volgens de richtlijnen twee verbaal hoofden en twee subjecten getagd (mag dat wel? is meer dan een *hd* niet tegen de principes van Alpino?). In xpath:

```
node[count(./node[@rel='su']) =2 and count(./node[@rel='hd']) =2]
```

Dit vindt echter niets. Alpino geeft voor het tweede voorbeeld een analyse met dp's erin:

```
//node[following-sibling::node/node[@rel="su"]/@lemma=./node[@rel='su']/@lemma and following
```

Of eigenlijk preciezer

```
//node[following-sibling::node/node[@rel="su"] [preceding-sibling::node[@rel='hd']] /@word=.
```

Helaas alleen voorbeelden met *zeggen* gevonden.

6.4 Apokoinouconstructies

Een apokoinouconstructie is een constructie waarbij een woord of woorden-groep tegelijkertijd deel uitmaakt, eerst als staart en dan als kop, van twee onafhankelijke constructies.

- maar nu [hadden we hier zo {*de vaart*} , *noemen wij dat*].

In het GCND wordt hier het categorielabel *apokoinou* gebruikt, en in het tweede zinsdeel is er een lege knoop die met het woord dat twee rollen heeft gecoïndiceerd is (=dezelfde index heeft).

```
//node[@cat="apokoinou"]
```

6.5 Opsomming van cijfers met betekenis ‘ongeveer’

Type 1

- gastjes van vijf zes jaar

```
//node[@cat='conj'] [count(./node[@pt='tw']) > 1 and count(./node[@pt="tw"]) = count(./node
```

Type 2

- een boek of twee drie

De een ... of determiner is te vinden met

- en k · zegge k · e ik nog een **een** stinkebol **of twee**.

```
//node[@cat='detp'
and node [@rel='mod' and @cat='mwu'] [node[@rel='mwp' and @pt='lid'] and node[@rel='mwp' and
]
```

```
//node[@rel='det' and
node[@cat='detp' and @rel="cnj"
and node [@rel='mod' and @cat='mwu'] [node[@rel='mwp' and @pt='lid'] and node[@rel='m
```

]]

6.6 Discontinue telwoorden drie jaar en half, drie kilo een half

- k · ware nog een jaar en half te jong.

Volgens de annotatierichtlijnen getagds als mwu met mwp delen.

- en da was maar op twee uren en half .

```
//node[@cat="mwu"] [node[@lemma="half"] [../node[@pt="tw"]/@begin < @begin]]
```

In de praktijk gebeurt het vaak anders:

- als mijn va . . . awel als mijn vader zijn zuster zes jaar en half was ging ze al naar het fabriek .

6.7 Tijdsbepalingen met (van) te: *te zessen, van te zessen*

- en het was dan al rond te vieren .

```
//node[node[@lemma='te' and @rel='hd'] and node[@pt='tw' and @rel='obj1' ]]
```

6.8 De die

- awel de die was heel klein hé .

```
//node[  
node[@rel='det' and @pt="lid"] and  
node[@rel="hd" and @pt="vnw" and @vwtype="aanw"]  
]
```

6.9 Code-switches naar het Frans/Engels/...

- ja de potten waren à peu près ten einde dan .

```
//node[@cat='mwu'] [node[@postag='SPEC(vreemd)'] and not (node[@postag!='SPEC(vreemd)'])]
```

6.10 Geluiden en klanknabootsingen

- a. en als je voeten zweetten **zwiep** zat je kleine teen erdoor.
- b. zodus iedere keer dat hij sloeg hé dat was . . . **djoef**.

Worden getagd als *tsw*, al dan niet met in zinsverband geannoteerde syntactische functie (*predc* bij b.)

```
//node[@pt='tsw']
```

Natuurlijk zijn lang niet alle tussenwerpsels klanknabootsingen.

6.11 Exclamatieve infinitiefzinnen

Zinnen zoals

- a. wij maar werken!
- b. en ik zoeken maar!

worden in het GCND als infinitieven (categorielabel *inf*) geanalyseerd, maar mét een overt subject (*su*). Deze infinitief wordt i.p.v. een smain gebruikt.

Het moet eenvoudiger kunnen, maar hieronder een benadering:

```
//node[@cat="inf" and not (@rel="vc")][
  node[@wvorm='inf' and @rel='hd'] and node[@rel='su'][descendant-or-self::node[@word]]
]
[not (descendant::node[@pt='ww' and @wvorm != 'inf'])]
[not (ancestor::node[@cat="whq" or @cat="oti" or @cat="ti" or @cat="smain"])]
[count(descendant::node[@pt="ww"]) = 1]
```