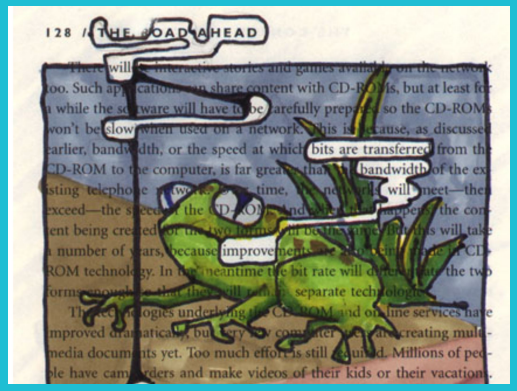
/instituut voor
de Nederlandse
taal/

Out of the Frying-Pan into the Fire

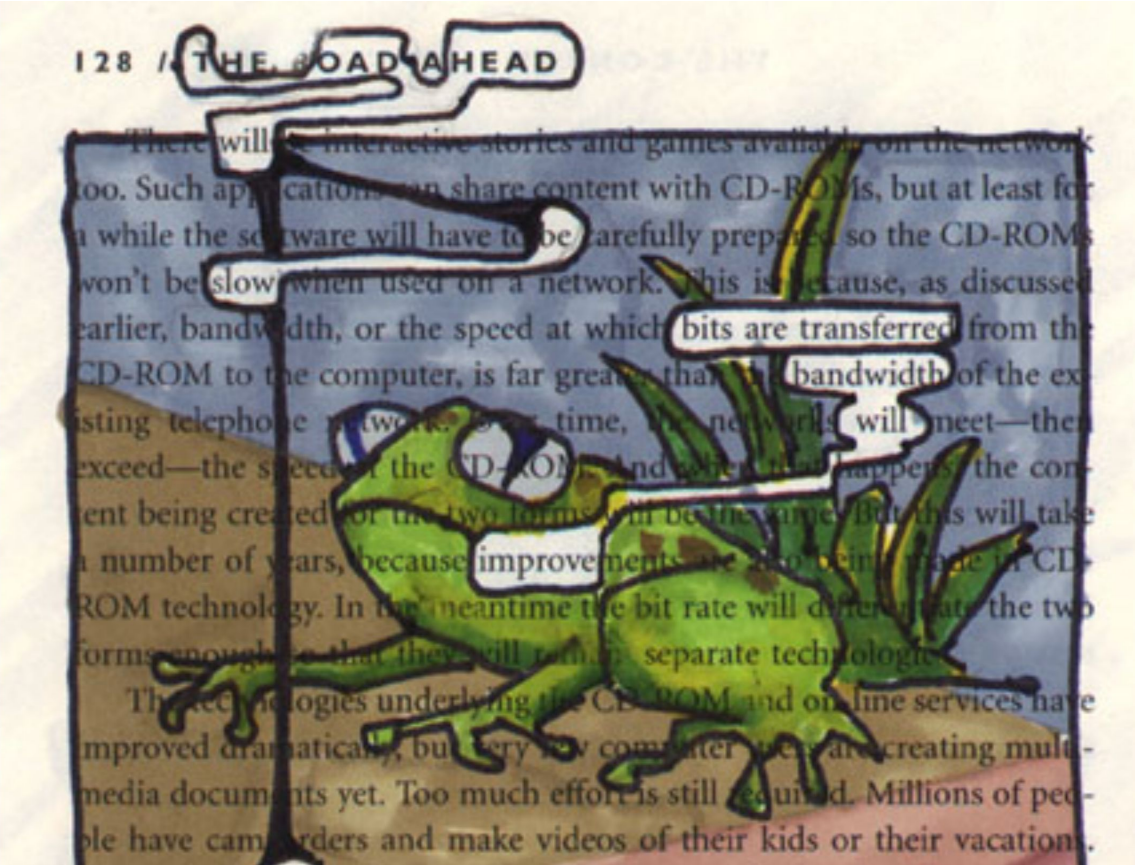
corpus annotation



Katrien Depuydt, Jesse de Does
katrien.depuydt@ivdnt.org
jesse.dedoes@ivdnt.org

Out of the Frying-Pan into the Fire

corpus annotation



Corpus annotation layers

- *Basic layers*:
  - lemma
  - part of speech
  - named entities
    - recognition
    - linking
- *Syntax*: treebanks
- *Semantics/pragmatics:*
  - word sense annotation
  - coreference

Annotation tools

- PoS taggers (+ lemmatizers), NER, parsing, semantic tagging, …
  - Frog https://languagemachines.github.io/frog/ (online version in TTNWW)
  - OpenConvert http://openconvert.clarin.inl.nl (modern Dutch)
  - "INL labs" http://inl-labs.inl.nl (historical dutch)
- Run on own hardware / Web-service based

Input and output

- Input: usually plain text or some XML
  - TEI http://www.tei-c.org
  - FoLiA https://proycon.github.io/folia/
- Output: (verbose) XML or plain text

```
1       Marie   Marie   [Marie] SPEC(deeleigen)                 1.000000        B-PER   B-NP    2       su
2       vroeg   vragen  [vraag] WW(pv,verl,ev)                  0.532544        0       B-VP    0       ROOT
3       zich    zich    [zich]  VNW(refl,pron,obl,red,3,getal)  0.999740        0       B-NP    2       se
4       af      af      [af]    VZ(fin)                         0.996853        0       0       2       svp
5       of      of      [of]    VG(onder)                       0.733333        0       B-SBAR  4       vc
6       hij     hij     [hij]   VNW(pers,pron,nomin,vol,3,ev,masc)      0.999659        0       B-NP    8       su
7       nog     nog     [nog]   BW()                            0.999930        0       B-ADVP  8       mod
8       zou     zullen  [zal]   WW(pv,verl,ev)                 0.999947        0       B-VP    5       body
9       komen   komen   [kom][en]       WW(inf,vrij,zonder)     0.861549        0       I-VP    8       vc
10      .       .       [.]     LET()                           0.999956        0       0       9       punct
```

example PoS tagging: modern

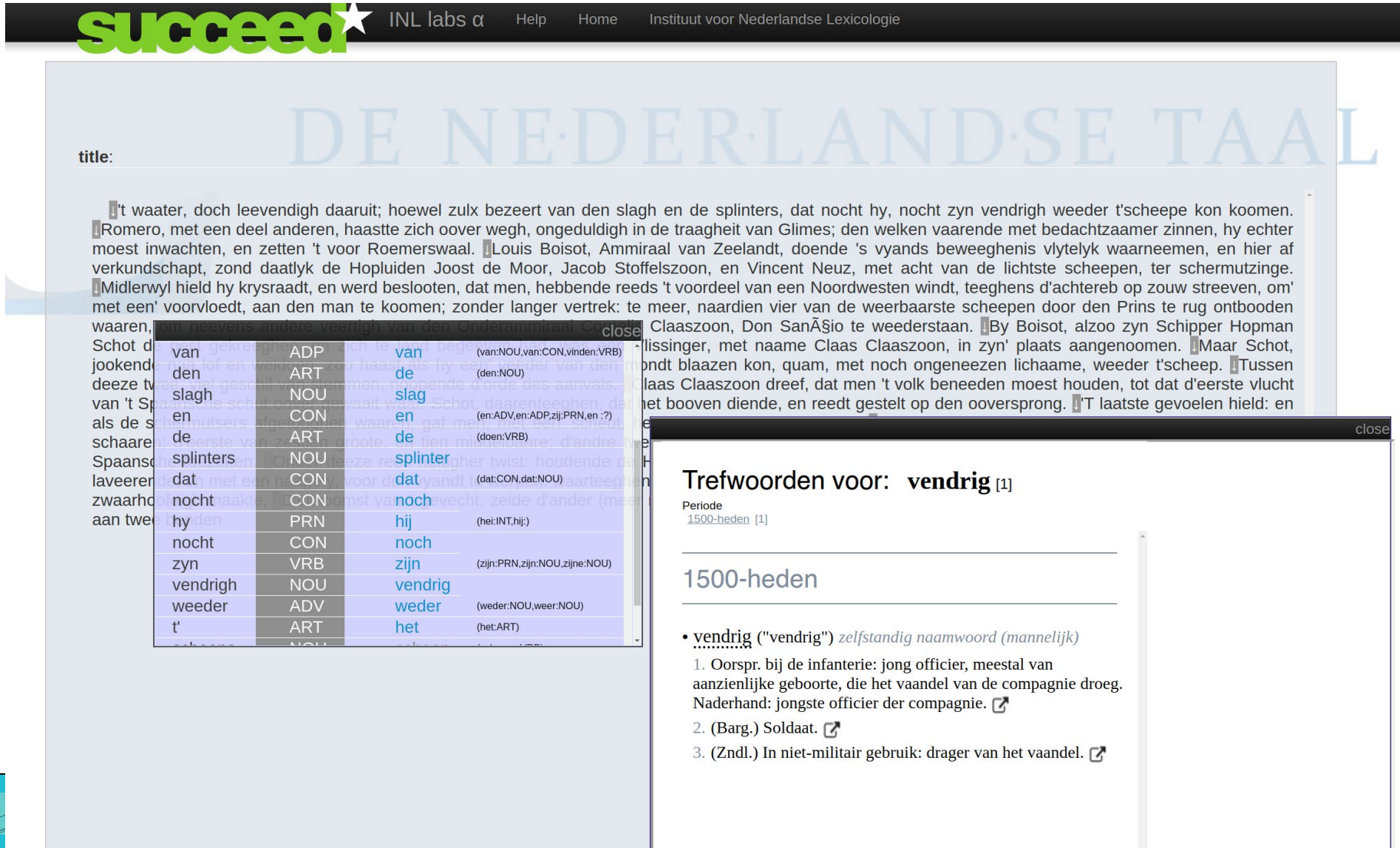
```
<s>
 <w lemma="de" type="PD(type=d-p,subtype=art-def)" xml:id="w.0">De</w>
 <w lemma="tovenaar" type="NOU-C(gender=f|m,number=sg)" xml:id="w.1">tovenaar</w>
 <w lemma="en" type="CONJ(type=coor)" xml:id="w.2">en</w>
 <w lemma="de" type="PD(type=d-p,subtype=art-def)" xml:id="w.3">de</w>
 <w lemma="hobbit" type="NOU-C(gender=f|m,number=sg)" xml:id="w.4">hobbit</w>
 <w lemma="duwen" type="VRB(finiteness=fin,mood=ind,tense=past,number=pl)" xml:id="w.5">duwden</w>
 <w lemma="het" type="PD(type=d-p,subtype=art-def)" xml:id="w.6">het</w>
 <w lemma="zwaar" type="AA(degree=pos,position=prenom,formal=infl-e)" xml:id="w.7">zware</w>
 <w lemma="krakend" type="AA(degree=pos,position=prenom,formal=infl-e)" xml:id="w.8">krakende</w>
 <w lemma="hek" type="NOU-C(gender=n,number=sg)" xml:id="
 <w lemma="open" type="AA(degree=pos,position=adv|pred)" xn
 <w lemma="en" type="CONJ(type=coor)" xml:id="w.11">en</w
 <w lemma="lopen" type="VRB(finiteness=fin,mood=ind,tense=pa
 <w lemma="over" type="ADP(type=pre)" xml:id="w.13">over</
 <w lemma="een" type="PD(type=indef,subtype=art-indef)" xml:i
 <w lemma="breed" type="AA(degree=pos,position=prenom)" xm
 <w lemma="pad" type="NOU-C(gender=n,number=sg)" xml:id="
 <w lemma="naar" type="ADP(type=pre)" xml:id="w.17">naar</
 <w lemma="het" type="PD(type=d-p,subtype=art-def)" xml:id="
 <w lemma="huis" type="NOU-C(gender=n,number=sg)" xml:id=
     <lb />
 </w>
 <pc type="post" xml:id="pc.000001">.</pc>

</s>
```

De tovenaar en de hobbit duwden het zware krakende hek open en liepen over een breed pa

close

De	PD(type=d-p,subtype=art-def)	de
tovenaar	NOU-C(gender=f|m,number=sg)	tovenaar
en	CONJ(type=coor)	en
de	PD(type=d-p,subtype=art-def)	de
hobbit	NOU-C(gender=f|m,number=sg)	hobbit
duwden	VRB(finiteness=fin,mood=ind,tense=past,number=pl)	duwen
het	PD(type=d-p,subtype=art-def)	het
zware	AA(degree=pos,position=prenom,formal=infl-e)	zwaar
krakende	AA(degree=pos,position=prenom,formal=infl-e)	krakend
hek	NOU-C(gender=n,number=sg)	hek
open	AA(degree=pos,position=adv|pred)	open
en	CONJ(type=coor)	en
liepen	VRB(finiteness=fin,mood=ind,tense=past,number=pl)	lopen
over	ADP(type=pre)	over

Part of speech tagging: historical



Example text

en den heer Direkteur
van het Bederlaers Werkhuis
Brugge.

Bruggen den 27 Maert 1864

Mynheer
Den ondergeteekende op gesloten in het bedelaers-werkhuis van Brugge sedert den 28 Oogst 1862 heeft de eer
UEu edele te verzoeken
my te willen voorstellen aen den heer Burgmeester van Bruggen voor het bekomen myner vryheid. *Zynde nu juist in het goed voor hetgeen mynen stiel betreft ik durf my op uwe gewoone welwillend heid betrouwen om eene goede antwoord op mynen brief te bekomen In afwachting blyve ik met den grootsten eerbied*

Uwen onderdanigen Dienaer
Franciscus Hendrickx

```
Zynde   zynde   N(soort,ev,basis,zijd,stan)
nu      nu      BW()
juist   juist   BW()
in      in      VZ(init)
het     het     LID(bep,stan,evon)
goed    goed    ADJ(vrij,basis,zonder)
voor    voor    VZ(init)
hetgeen hetgeen VNW(betr,det,stan,nom,zonder,zonder-n)
mynen   mynen   N(soort,mv,basis)
stiel   stiel   N(soort,ev,basis,zijd,stan)
betreft betreffen       WW(pv,tgw,met-t)
ik      ik      VNW(pers,pron,nomin,vol,1,ev)
durf    durven  WW(pv,tgw,ev)
my      my      SPEC(vreemd)
op      op      VZ(init)
uwe     uw      VNW(bez,det,dial)
gewoone gewoone N(soort,ev,basis,zijd,stan)
welwillend      welwillend      ADJ(vrij,basis,zonder)
heid    heid    WW(od,vrij,zonder)
betrouwen       betrouwen       WW(inf,vrij,zonder)
om      om      VZ(init)
eene    eene    ADJ(prenom,basis,met-e,stan)
goede   goed    ADJ(prenom,basis,met-e,stan)
antwoord        antwoord        N(soort,ev,basis,onz,stan)
op      op      VZ(init)
mynen   mynen   N(soort,mv,basis)
brief   brief   N(soort,ev,basis,zijd,stan)
te      te      VZ(init)
bekomen bekomen WW(inf,vrij,zonder)
In      in      VZ(init)
afwachting      afwachting      N(soort,ev,basis,zijd,stan)
blyve   blyve   WW(pv,verl,ev)
ik      ik      VNW(pers,pron,nomin,vol,1,ev)
met     met     VZ(init)
den     de      LID(bep,dial)
grootsten       grootsten       ADJ(dial)
eerbied eerbied N(soort,ev,basis,zijd,stan)
```

```
Zynde   VRB     zijn
nu      ADV     nu
juist   ADV     juist
in      ADP     in
het     ART     het
goed    NOU     goed
voor    ADP     voor
hetgeen PRN     hetgeen
mynen   PRN     mijn
stiel   NOU     stiel
betreft VRB     betreffen
ik      PRN     ik
durf    VRB     durven
my      PRN     ik
op      ADP     op
uwe     PRN     uw
gewoone ADJ     gewoon
welwillend      NOU     welwillend
heid    NOU     heide
betrouwen       VRB     betrouwen
om      CON     om
eene    ART     een
goede   ADJ     goed
antwoord        NOU     antwoord
op      ADP     op
mynen   PRN     mijn
brief   NOU     brief
te      ADP     te
bekomen VRB     bekomen
In      ADP     in
afwachting      NOU     afwachting
blyve   VRB     blijven
ik      PRN     ik
met     ADP     met
den     ART     de
grootsten       ADJ     groot
eerbied NOU     eerbied
```

Tagger invocation example

```
import scala.xml._

object Tagger
{
  def bab_tagger_url(text:String):String =
   s"http://inl-labs.inl.nl/succeed/text?input=${java.net.URLEncoder.encode(text)}&format=text&tagger=bab-tagger&output=raw"
  def chn_tagger_url(text:String):String =
   s"http://openconvert.clarin.inl.nl/openconvert/text?input=${java.net.URLEncoder.encode(text)}&format=text&tagger=chn-tagger&output=raw&to=tei"

  def getTextButNotIn(x:Node,tagName:String):String =
     if (x.isInstanceOf[Text])
       x.text
     else
       x.child.filter(_.label != tagName).map(getTextButNotIn(_,tagName)).mkString("")

  def tag(f:String=>String, text:String) =
  {
    val d = XML.load(f(text))
    val z = (d \\ "w").map(w =>
      List(
         ("word", getTextButNotIn(w,"interp")),
         ("lemma", (w \ "@lemma").text),
         ("pos", (w \ "@type").text)))
    z.foreach(println)
  }
}

Tagger.tag(Tagger.bab_tagger_url,"De Aerde nu was woest ende ledich, ende duysternisse was op den afgront: ende de Geest Godts sweefde op de
    Wateren.")
```

Named entity recognition: “modern”

Die dag zongen ze niet en vertelden ook geen verhalen, hoewel het weer beter werd; en ook de volgende dag en de dag daarna niet. Ze voelden nu dat het gevaar aan weerskanten niet ver weg was. Zij overnachtten onder de sterren en hun paarden kregen meer te eten dan zij, want er was volop gras, maar er zat niet veel in hun zakken, ook niet met dat wat zij van de trollen hadden genomen. Op een morgen doorwaadden zij de rivier op een brede, ondiepe plek vervuld van het lawaai van ratelende stenen en brui - send schuim. De andere oever was steil en glibberig. Toen zij de top ervan hadden bereikt, hun pony's leidend, zagen zij dat zij de grote bergen heel dicht genaderd waren. Zij schenen nu nog slechts een dag gaans van de voet van de dichtstbijzijnde verwijderd te zijn. Hij zag er donker en naargeestig uit, hoewel er vlekken zonlicht op de bruine hellingen lagen, en achter zijn rondingen glinsterden de toppen van sneeuwkappen. 'Is dat De Berg?' vroeg Bilbo op plechtige toon, terwijl hij er met ronde ogen naar keek. Hij had nog nooit eerder zo iets groots gezien.

'Natuurlijk niet!' zei Balin. Mat is slechts het begin van de Nevelbergen en wij moeten er op de een of andere manier door, of erover, of eronder voor - dat wij Wilderland erachter kunnen bereiken. En het is ook van de andere kant nog een behoorlijk eind naar de Eenzame Berg in het oosten waar Smaug op onze schat ligt.

'0!' zei Bilbo en op dat ogenblik voelde hij zich vermoeider dan hij zich ooit had gevoeld. Hij moest weer denken aan zijn luie stoel voor de haard in de geliefde zitkamer van zijn hobbit-hol en aan de ketel die zong. En niet voor de laatste keer!

Nu ging Gandalf voorop. 'Wij mogen de weg niet mislopen, anders zijn we er geweest,' zei hij. 'Wij hebben in de eerste plaats eten nodig, en rust in re - delijke veiligheid - ook is het absoluut noodzakelijk om de Nevelbergen langs het goede pad in te trekken, anders raak je erin verdwaald en moet je teruggaan en weer helemaal opnieuw beginnen (als je tenminste nog te - rugkomt).' Ze vroegen hem waar hij heenging en hij antwoordde: 'Jullie zijn nu vlak bij de rand van de Wildernis gekomen, zoals sommigen van jullie wellicht weten. Ergens voor ons, verscholen, ligt het schone dal Ri - vendell, waar Elrond woont in het Laatste Huiselijke Huis. k heb mijn vrienden een boodschap meegegeven en wij worden verwacht.'

Namescape: NER for literary studies

tagger	NE type	precision	recall	F1
Stanford	location	0.802	0.712	0.754
	misc	0	0	0
	organisation	0.433	0.228	0.299
	person	0.876	0.895	0.881
	overall	0.853	0.824	0.838
Namescape	location	0.83	0.729	0.776
	misc	0	0	0
	organisation	0.516	0.251	0.339
	person	0.867	0.917	0.896
	overall	0.853	0.838	0.845
Namescape (+wv)	location	0.858	0.830	0.844
	misc	0.4	0.154	0.222
	organisation	0.656	0.459	0.54
	person	0.932	0.941	0.936
	overall	0.904	0.881	0.893

Named entity recognition: historical

INT sequence taggers

Taggers

- PoS (modern/historical)
- NER

Technical details

- SVM*light* (c++) for training
- Tagging in java
- Historical lexicon used for lemmatization, in combination with approximate match
- TEI input and output possible
- Uses word embeddings and (for NER) other semi-supervised features
- Currently published only as web services
  - training procedure invokes executable for SVM-light, rather involved

But….

This is all very nice but:

- Tools perform badly for non-standard-modern language
- Developing training material for other material is a huge effort

Required effort: training/eval data

To do: harmonize,share, make benchmark sets

Corpus Gysseling	1.6M
CRM	650k
Adelheid training sets	?
Gentse Brieven	70k
Brieven als Buit	400k
Hooft letters	?
Statenvertaling	?
“Corpus Vosters”	~100k
IMPACT evaluation corpora	~20k
Nederlab evaluation data (tbd)	At least 10k per century

Tagset issues

CGN/DCOI dominant for modern Dutch, but some issues with historical Dutch:

- Principle *tag + lemma ⇒ (almost) determines word form* cannot be maintained historically
- (likewise for CG/CRM): tagging inflection instead of case/number/mood/tense etc does not benefit (diachronic) analysis
- Details:
  - Mood cannot be lumped with tense historically (PVTIJD = tegenwoordig, verleden, imperatief, conjunctief.)
  - "dial" for all non-standard-modern words is unsatisfactory

Diachronic and synchronic tagsets

- CGN/DCOI http://www.hum.uu.nl/medewerkers/p.monachesi/papers/vaneynde.pdf
- Corpus Gysseling/Corpus van Reenen-Mulder
- Adelheid tagset http://adelheid.ruhosting.nl/download/Adelheid1_TagsetManual.pdf
- Main PoS only in Brieven als Buit en Gentse Spelen
- GiGaNT tagset
  http://www.ivdnt.org/images/stories/onderzoek_en_onderwijs/publicaties/TaalbankWorkingpaper3.pdf
- International standards (Gold, eagles, parole, …)
  - Universal dependencies, adopted in CLARIN/CLARIAH
    http://universaldependencies.org/

Smarter software?

Can we avoid a huge effort of training corpus building??

- More advanced supervised learning (deep….)
- Domain adaptation techniques
- Semi-supervised techniques
- Map to “modernized” Dutch
- Derive lexica from parallel text
- Transfer tagging by alignment

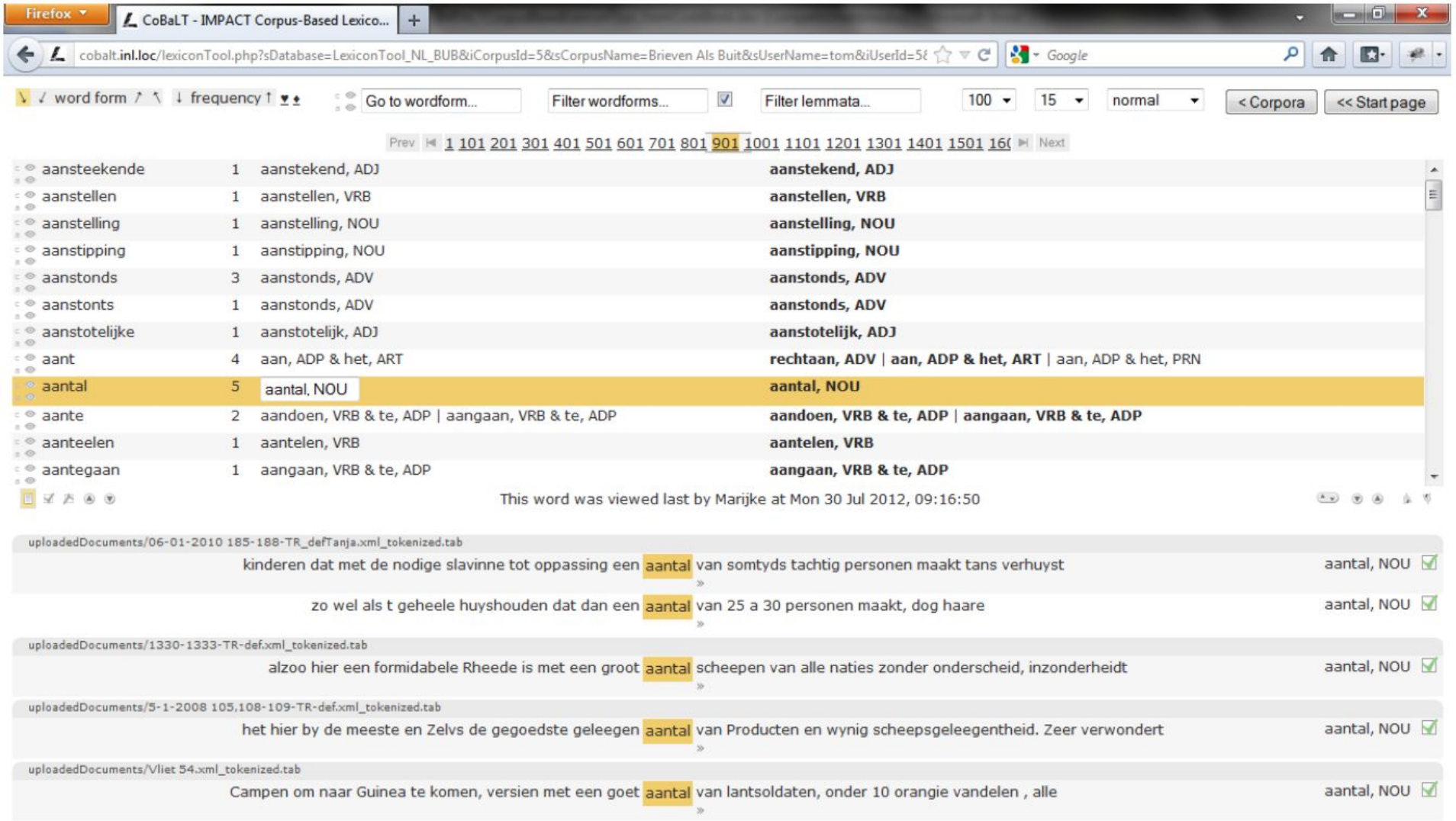
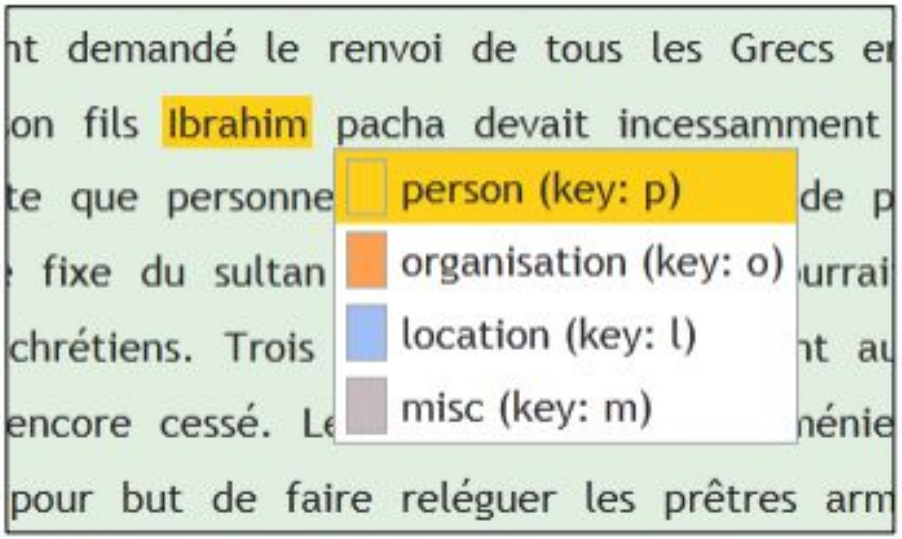
Tools for manual annotation

https://github.com/INL/COBALT

CoBaLT is an application in which a corpus of texts can be loaded so as to be able to annotate its tokens (lemmatize and more). The annotation work in CoBaLT gives two products: an annotated corpus, and a lexicon consisting of word forms and their corresponding lemmata and such assigned to them.

https://github.com/INL/AttestationTool

The Attestation Tool is a multi purpose GUI used in the production of computational lexica and gold standard data for NE tagging. It addresses the situation of having a comprehensive dictionary consisting of headwords (lemmata) and corresponding quotations in which occurrences of the headwords were automatically matched.


Firefox
CoBaLT - IMPACT Corpus-Based Lexico...
cobalt.inl.loc/lexiconTool.php?sDatabase=LexiconTool_NL_BUB&iCorpusId=5&sCorpusName=Brieven Als Buit&sUserName=tom&iUserId=5
Google
word form
frequency
Go to wordform...
Filter wordforms...
Filter lemmata...
100
15
normal
< Corpora
<< Start page
Prev 1 101 201 301 401 501 601 701 801 901 1001 1101 1201 1301 1401 1501 16(Next
aansteekende 1 aanstekend, ADJ aanstekend, ADJ
aanstellen 1 aanstellen, VRB aanstellen, VRB
aanstelling 1 aanstelling, NOU aanstelling, NOU
aanstipping 1 aanstipping, NOU aanstipping, NOU
aanstonds 3 aanstonds, ADV aanstonds, ADV
aanstonts 1 aanstonds, ADV aanstonds, ADV
aanstotelijke 1 aanstotelijk, ADJ aanstotelijk, ADJ
aant 4 aan, ADP & het, ART rechtaan, ADV | aan, ADP & het, ART | aan, ADP & het, PRN
aantal 5 aantal, NOU aantal, NOU
aante 2 aandoen, VRB & te, ADP | aangaan, VRB & te, ADP aandoen, VRB & te, ADP | aangaan, VRB & te, ADP
aanteelen 1 aantelen, VRB aantelen, VRB
aantegaan 1 aangaan, VRB & te, ADP aangaan, VRB & te, ADP
This word was viewed last by Marijke at Mon 30 Jul 2012, 09:16:50
uploadedDocuments/06-01-2010 185-188-TR_defTanja.xml_tokenized.tab
kinderen dat met de nodige slavinne tot oppassing een aantal van somtyds tachtig personen maakt tans verhuyst aantal, NOU
zo wel als t geheele huyshouden dat dan een aantal van 25 a 30 personen maakt, dog haare aantal, NOU
uploadedDocuments/1330-1333-TR-def.xml_tokenized.tab
alzoo hier een formidabele Rheede is met een groot aantal scheepen van alle naties zonder onderscheid, inzonderheidt aantal, NOU
uploadedDocuments/5-1-2008 105,108-109-TR-def.xml_tokenized.tab
het hier by de meeste en Zelvs de gegoedste geleegen aantal van Producten en wynig scheepsgeleegentheid. Zeer verwondert aantal, NOU
uploadedDocuments/Vliet 54.xml_tokenized.tab
Campen om naar Guinea te komen, versien met een goet aantal van lantsoldaten, onder 10 orangie vandelen , alle aantal, NOU

Named entity annotation tool=dictionary attestation tool