

Digging for domain-specific terms in North Saami

Ciprian-Virgil Gerstenberger
Berit Merete Nystad Eskonsipo
Márjá Eira

UiT The Arctic University of Norway

*INARI 2013
September 2013 Inari, Finland*



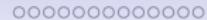
Motivation: Why?



Approach: How?



Results: What?



The **FAD** project

Fornyings-, administrasjons- og kirkedepartementet
Oðasmahttin-, hálddahus- ja girkodepartemeanta

Motivation: Why?

Approach: How?

Results: What?

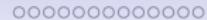
Motivation: Why?



Approach: How?



Results: What?



The **FAD** project

Fornyings-, administrasjons- og kirkedepartementet
Oðasmahttin-, hálddahus- ja girkodepartemeanta

Motivation: Why?

Approach: How?

Results: What?

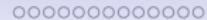
Motivation: Why?



Approach: How?



Results: What?



The **FAD** project

Fornyings-, administrasjons- og kirkedepartementet
Oðasmahttin-, hálddahus- ja girkodepartemeanta

Motivation: Why?

Approach: How?

Results: What?

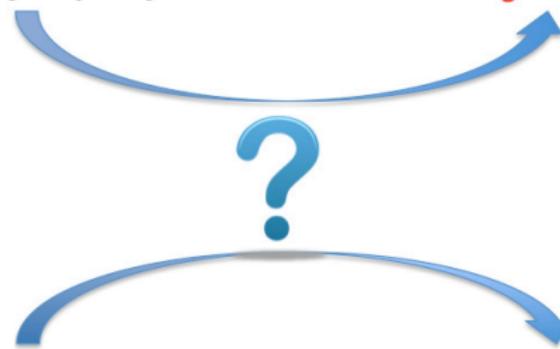
Are administrative North Saami terms extractable from the extant Norwegian-North Saami parallel corpus?

Jurddá *lámisoajuin lea...*

Jurddá *buhcciidpenšuvnnain lea...*

Hensikten med *uførepensjon* er...

Jurddá *bargonávccahisvuodaoajuin lea...*

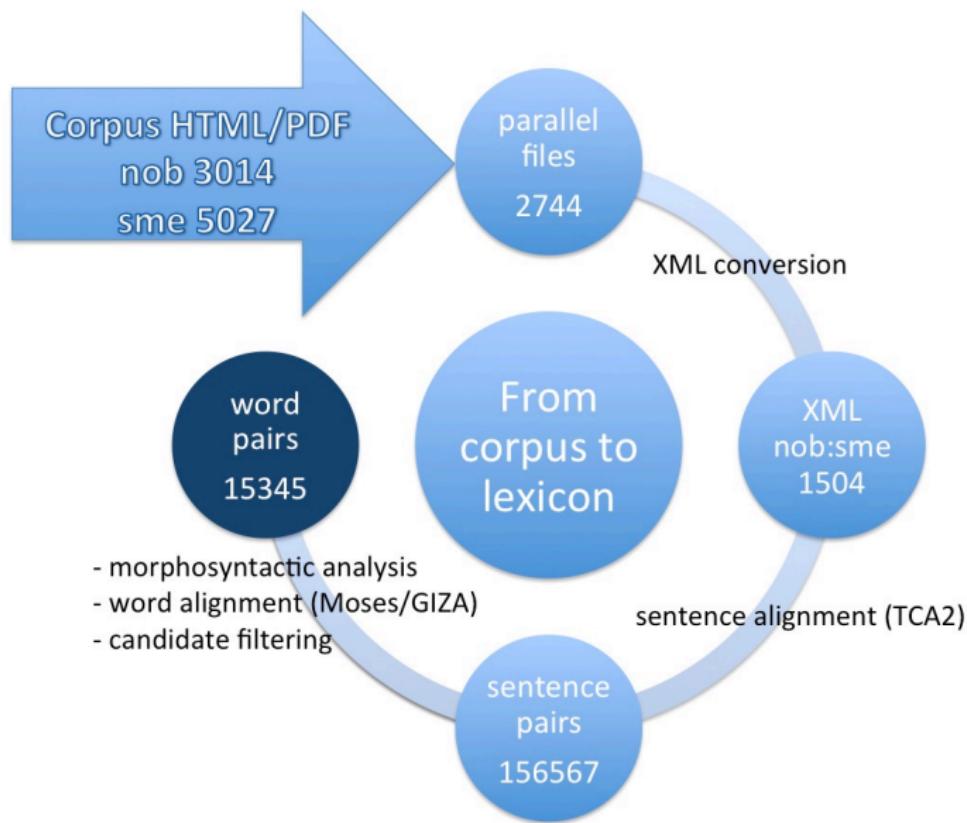


FORNYINGS-, ADMINISTRASJONS-
OG KIRKEDEPARTEMENTET



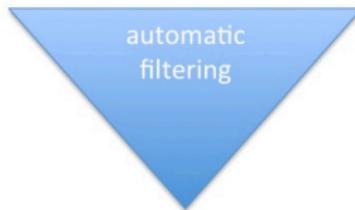
Sámi
giellatekno

Standard computational linguistic methods have been used.



A huge amount of manual work has been invested in picking up the correct wordaligned pairs.

```
1 0 -10.0 0.0 0.25 Vurderingsverktøyet<subst> árvvoštallanreaidu<N>
1 0 -10.0 0.0 0.0131579 Vurderingsverktøyet<subst> ovdánanságastallan<N>
1 0 -10.0 0.0 0.0084034 Vurderingsverktøyet<subst> gihpa<N>
1 0 -10.0 0.0 0.0011534 Vurderingsverktøyet<subst> sisdoallu<N>
```



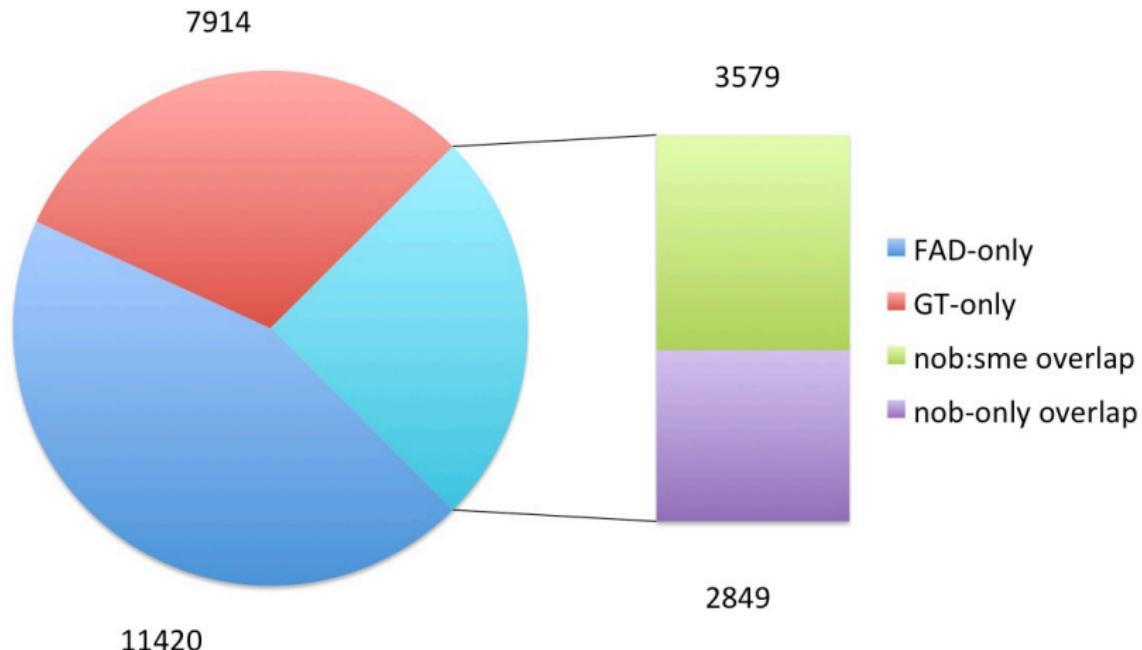
```
1 0 -10.0 0.0 0.25 Vurderingsverktøyet<subst> árvvoštallanreaidu<N>
```

```
1 0 -10.0 0.0 0.5 språklikestilling<subst> gielladásseárvu<N>
1 0 -10.0 0.0 0.2 kjønnslikestilling<subst> sohkabealdásseárvu<N>
1 0 -10.0 0.0 0.1666667 likestillingstest<subst> Fredrikke<N>
1 0 -10.0 0.0 0.1666667 likestillinganalyse<subst> dásseárvvoanaliiisa<N>
```



```
<e src="fad">
<lg>
  <l pos="N">språklikestilling</l>
</lg>
<mg>
  <tg xml:lang="sme">
    <t pos="N">gielladásseárvu</t>
  </tg>
</mg>
</e>
```

GT and FAD data have been compared and merged.



Meaning groups had to be manually unified.

```
<e>
  <lg>
    <l pos="N">barnevakt</l>
  </lg>
  <mg>
    <tg xml:lang="sme">
      <t pos="N" re="kvinnelig">mánnabiigd</t>
      <t pos="N" type="NomAg">mánnageahčíi</t>
      <t pos="N" type="NomAg">mánnóaidni</t>
    </tg>
  </mg>
</e>
```

(a) GT-only

```
<e src="fad">
  <lg>
    <l pos="N">likestillingsspørsmål</l>
  </lg>
  <mg>
    <tg xml:lang="sme">
      <t pos="N">dásseárvgogažaldat</t>
      <t pos="N">dásseárvoášši</t>
    </tg>
  </mg>
</e>
```

(b) FAD-only

```
<e>
  <lg>
    <l pos="N">valgforbund</l>
  </lg>
  <mg>
    <tg xml:lang="sme">
      <t pos="N" src="gt;fad">válgalihtru</t>
    </tg>
  </mg>
</e>
```

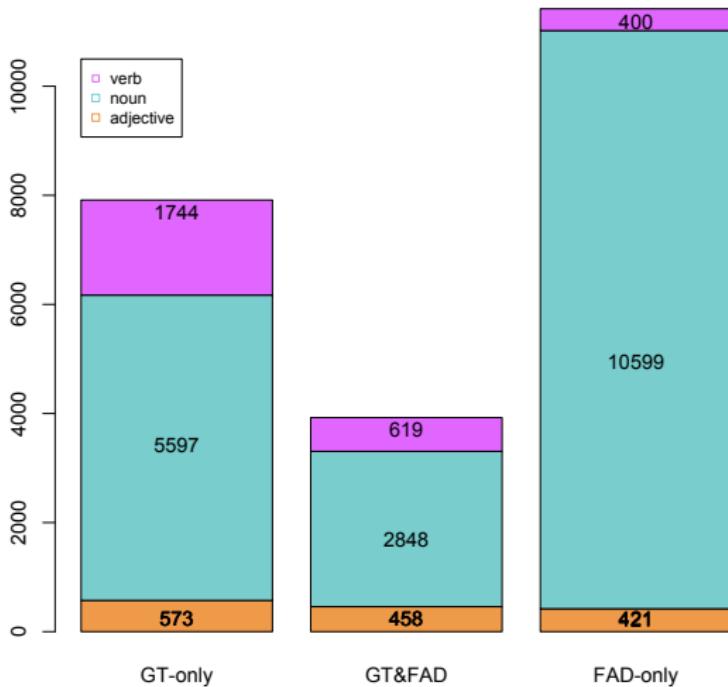
(c) GT&FAD total overlap

```
<e>
  <lg>
    <l pos="N">disposisjon</l>
  </lg>
  <mg>
    <tg xml:lang="sme">
      <t pos="N" src="gt;fad">disposišuvdna</t>
    </tg>
  </mg>
  <mg>
    <tg xml:lang="sme">
      <t pos="N" src="fad" re="jus">háldogeavaheapmi</t>
    </tg>
  </mg>
</e>
```

(d) GT&FAD partial overlap

In terms of word class, nouns constitute the biggest part.

nob lexical data



Giellatekno is **not** a language **normative** institution.



How to evaluate data?

Database lookup

```
<e src="fad">
  <lg>
    <l pos="N">likestillingsspørsmål</l>
  </lg>
  <mp>
    <tg xml:lang="sme">
      <t pos="N">dásseárvogažaldat</t>
      <t pos="N">dásseárvodáši</t>
    </tg>
  </mp>
</e>
```

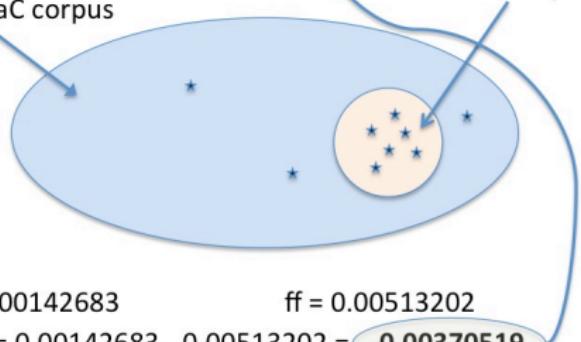


Term frequency

```
<e>
  <lg>
    <l pos="N" gf-ff="-0.00370519">disposisjon</l>
  </lg>
  <mp>
    <tg xml:lang="sme">
      <t pos="N" src="gt;fad" gf-ff="0.0001268">disposišuvdna</t>
    </tg>
  </mp>
  <mp>
    <tg xml:lang="sme">
      <re>jus</re>
      <t pos="N" src="fad" gf-ff="-0.00019637">háldogeavaheapmi</t>
    </tg>
  </mp>
</e>
```

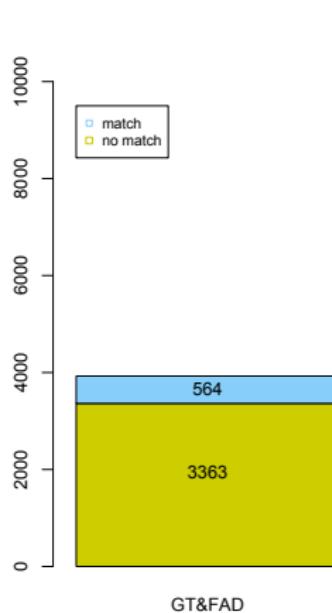
$g = \text{NoWaC corpus}$

$f = \text{FAD corpus}$



A lookup in a specialized termbase yields a total of about 1000 string matches.

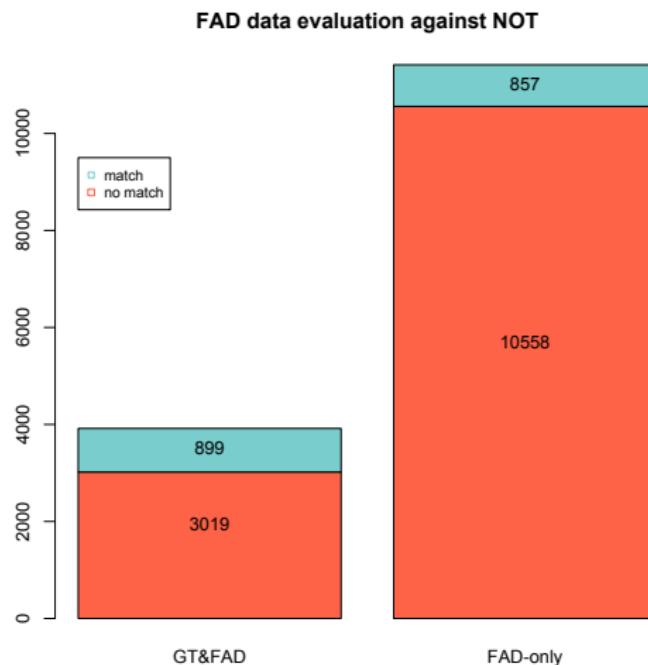
FAD data evaluation against KBN



KBN termbase

- 8473 terms
- business and administrative terminology
- bilingual
(English, Norwegian)

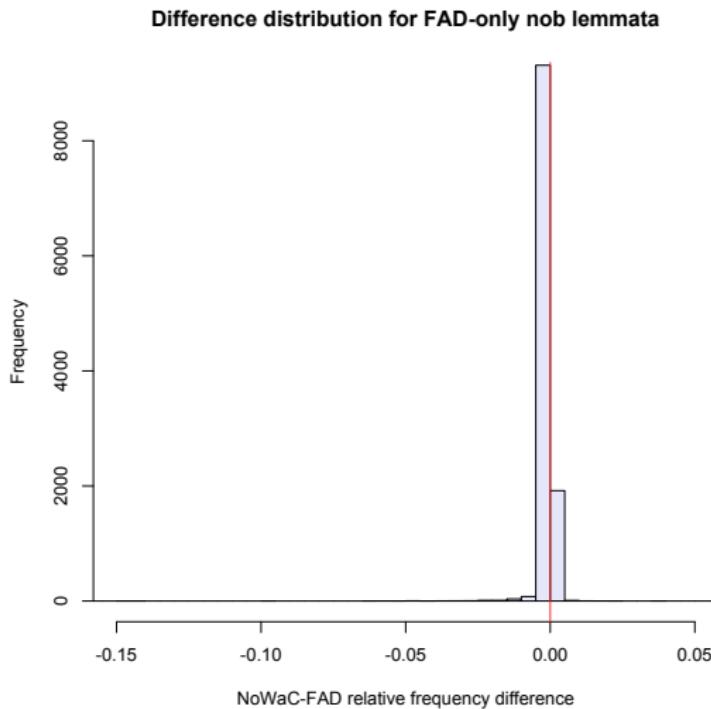
A string match of a nob lemma against a termbase is **no guarantee** for the sme translation to be a term.



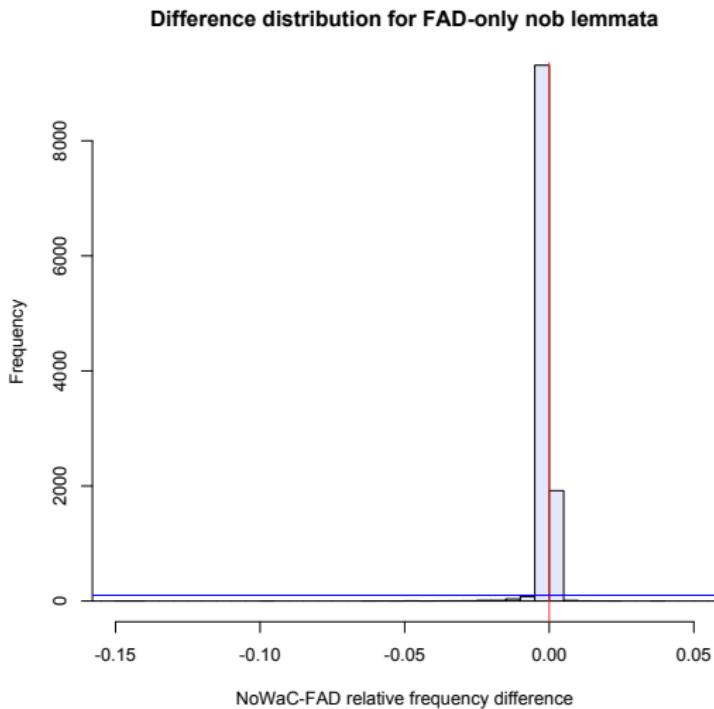
NOT termbase

- 30525 terms
- Norsk termbank
- multilingual
(English, Norwegian, etc.)

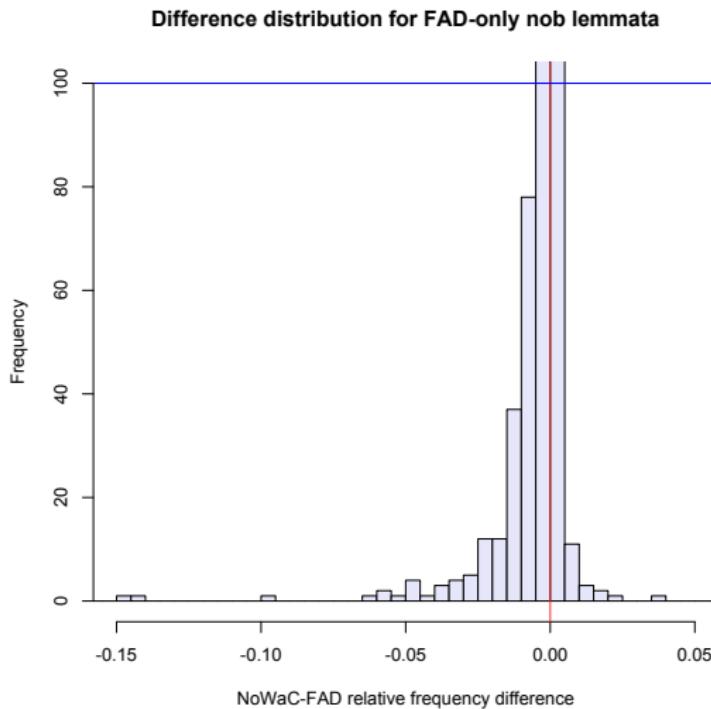
The FAD-only grouped frequency distribution shows a negative skew, more probable term candidates.



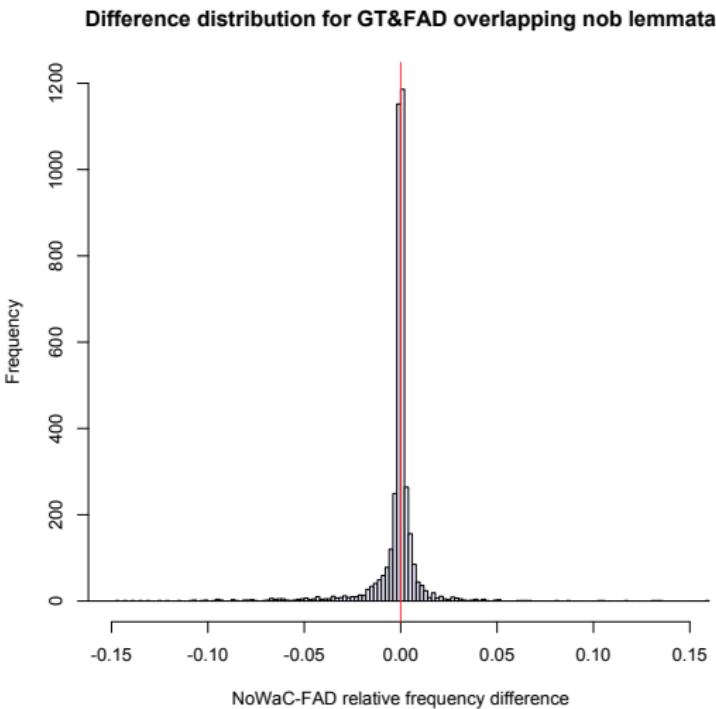
The FAD-only grouped frequency distribution shows a negative skew, more probable term candidates.



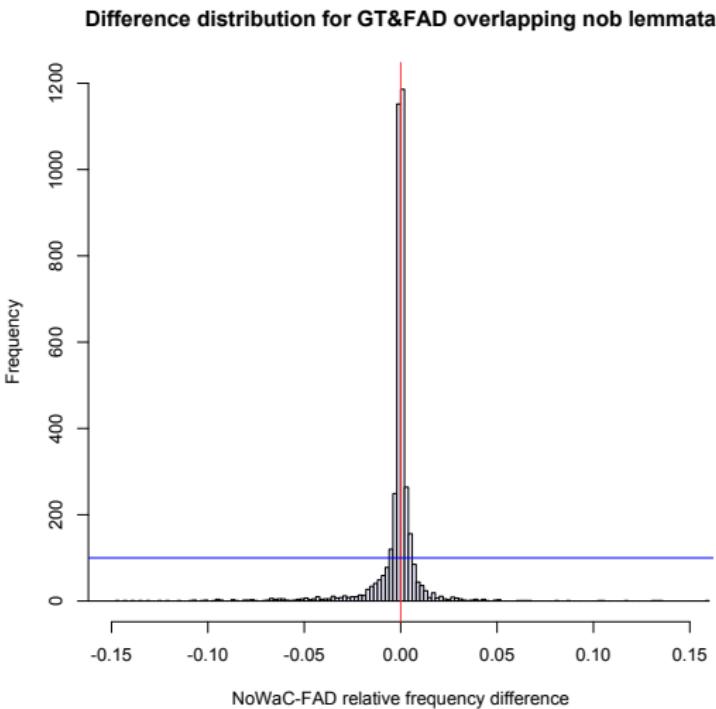
The FAD-only grouped frequency distribution shows a negative skew, more probable term candidates.



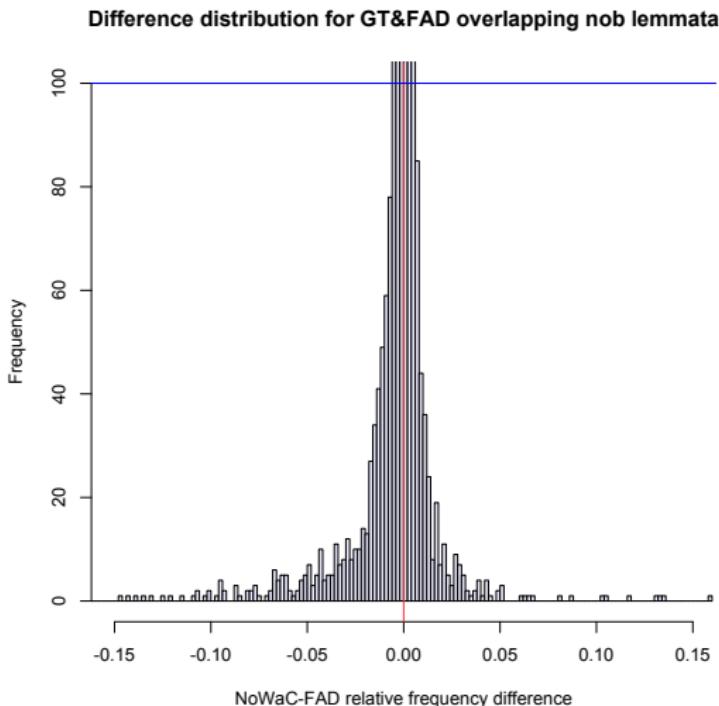
The GT&FAD grouped frequency distribution seems to be more balanced between negative and positive values.



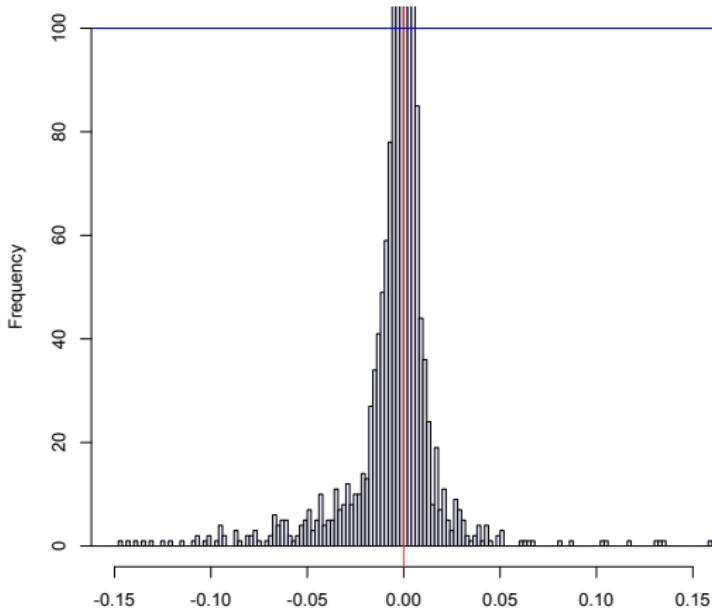
The GT&FAD grouped frequency distribution seems to be more balanced between negative and positive values.



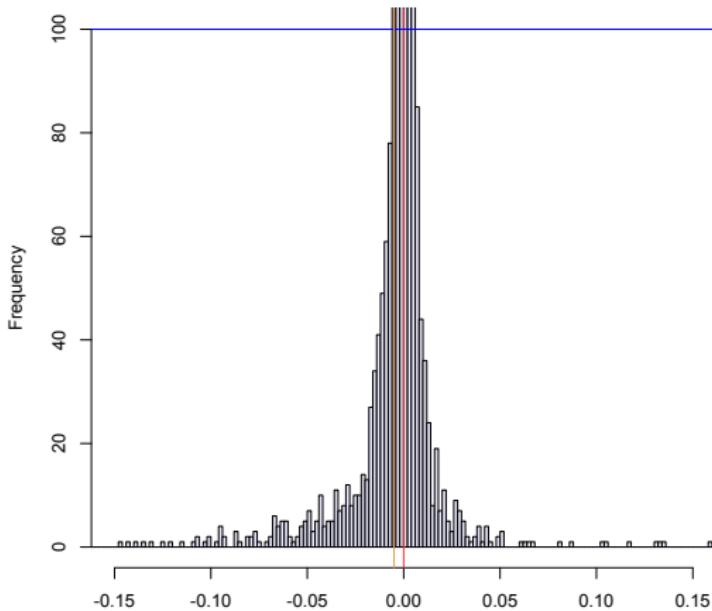
The GT&FAD grouped frequency distribution seems to be more balanced between negative and positive values.



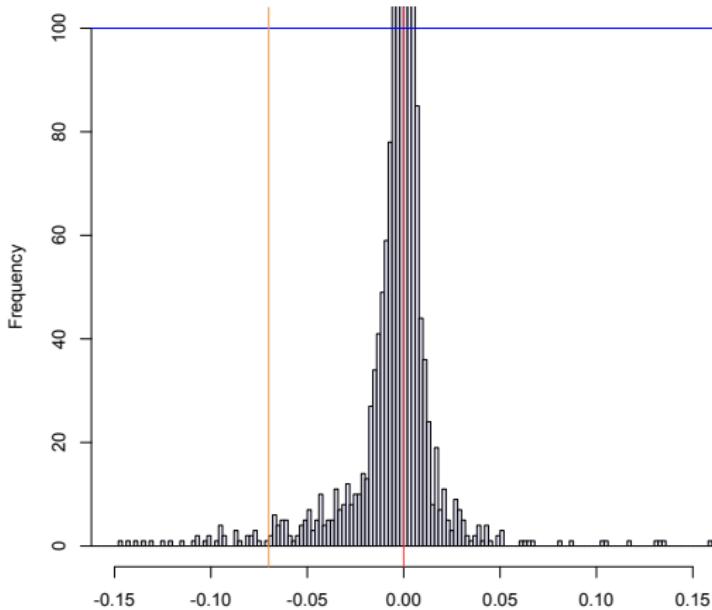
Is it possible to find a boundary between potential term candidates and ordinary words?



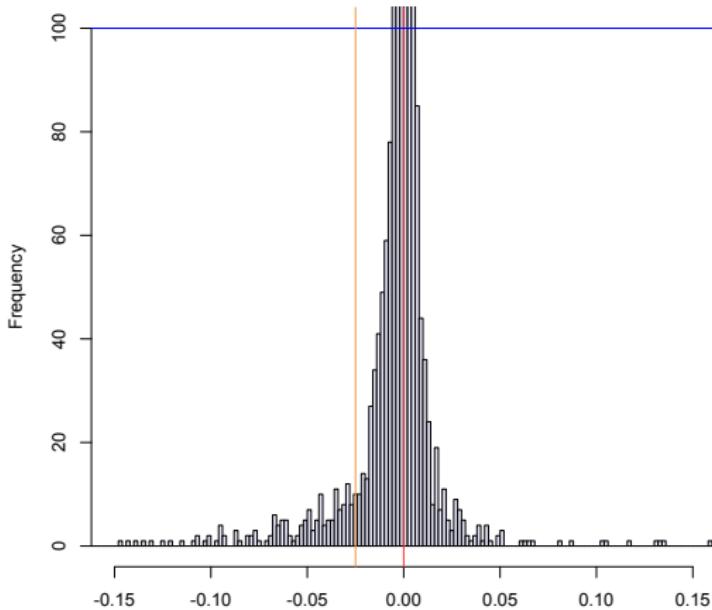
Is it possible to find a boundary between potential term candidates and ordinary words?



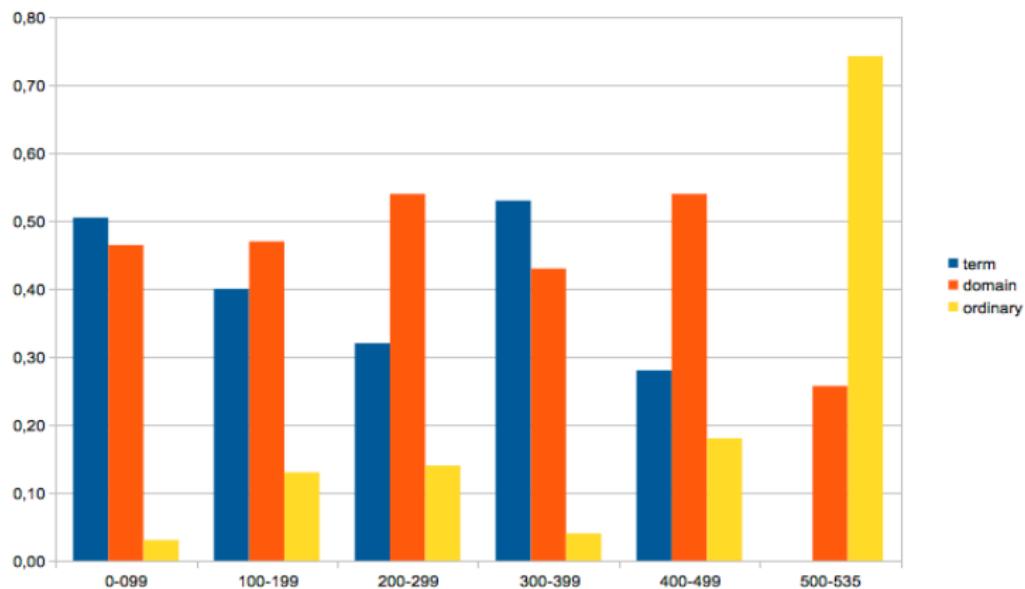
Is it possible to find a boundary between potential term candidates and ordinary words?

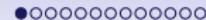
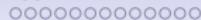


Is it possible to find a boundary between potential term candidates and ordinary words?



A manual evaluation of data sample showed no clear-cut termness delimitation.

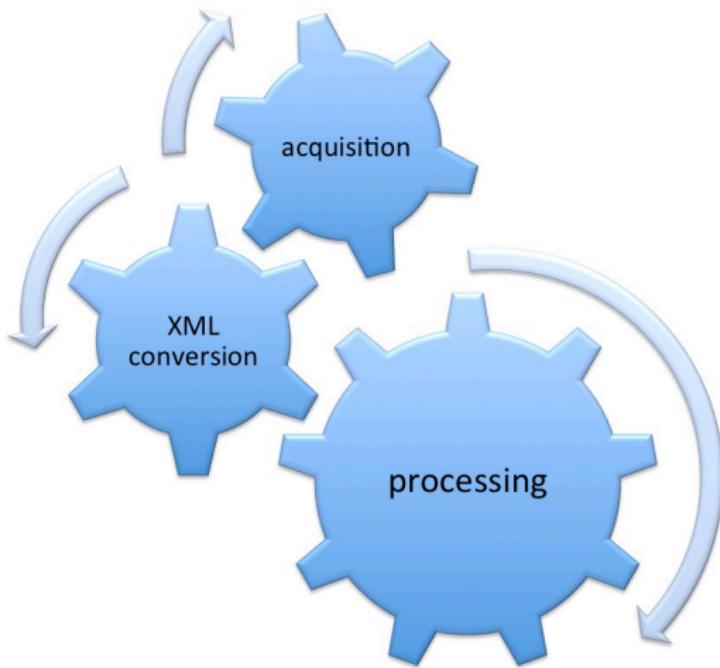


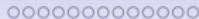


What are the results of the FAD project?

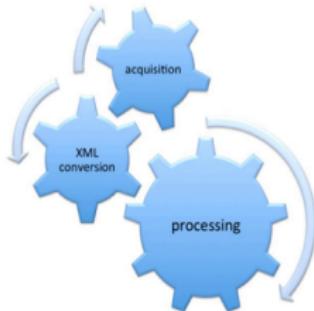


Corpus acquisition, conversion and testing routines have been greatly improved.

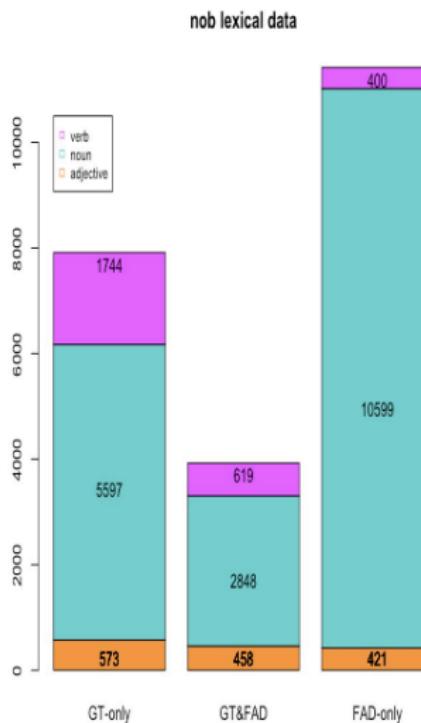




Corpus



Lexica and dictionaries have been tremendously extended.



Vuostai DigiSámit (2 funnel)

Alt **Vuostais Digráhku** Vuostais DigiSákooh Dictionary

hálddašanguovlu

zuhč
forvaltningsområde, administrativt område

Analyse: sg. nom.
Nakkelformer:

- sg. gen. hálddašanguovla
- sg. ill. hálddašanguovlu
- pl. ill. hálddašanguovluade

Risten.no version 2

All dictionary (96)
Norwegian words (3)
Samnordic terminology (1)
FAD-terms (164 words)

Risten

Ristenarbeider
Risten
Risten-tp
Filenesarbeider
Filenester
Risten
Risten-tele
Ristenkommune
Risten-tp
Ristenarbeid
Ristenarbeider

Náhddágisámit Ruskoo Lohkanvestiki Min tims

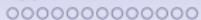
Dárgiella -- Davvísámegiella (1+ Muissa)

Dárgiella -- Davvísámegiella
Dárgiella -- Davvísámegiella
Dárgiella -- Suomagella
Suomagella -- Davvísámegiella

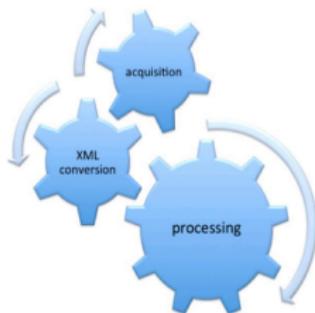
Earii sámmegjrit

Dárgiella -- Davvísámegiella (1+ Muissa)

språkforvaltningsområde (subst.)
1. gella/hálddašanguovlu
forvaltningsområde (subst.)
1. hálddašanguovlu (areal)
forvaltningsområdet för samisk språk
sámmegjrit hálddašanguovlu
2. hálddašanguorgi, hálddašanguovlu, hálddašánoessi,
hálddašanguovlu (bransje)
språk (subst.)
1. gella



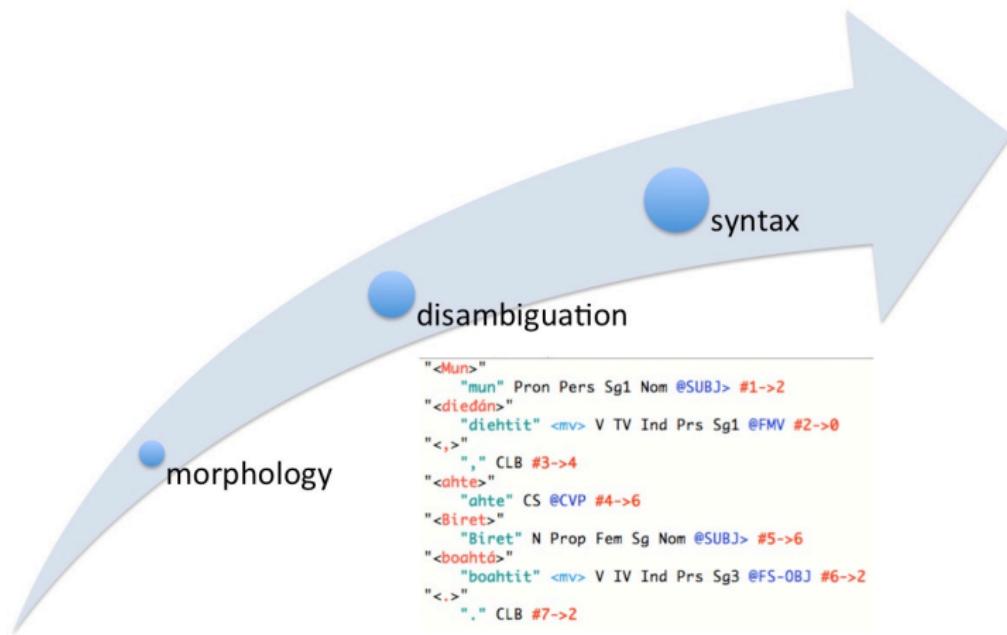
Corpus

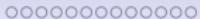


Lexicon

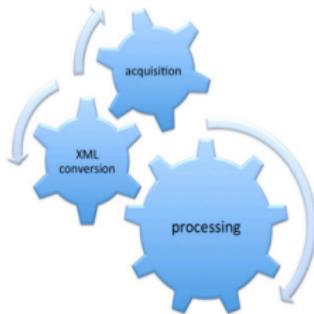


During the often repeated testing and correcting phases the linguistic analysis tools have been also steadily upgraded.

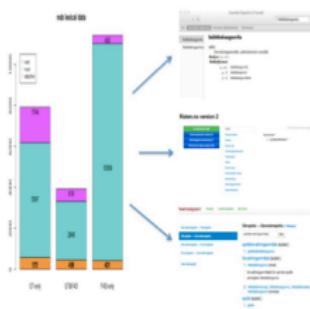




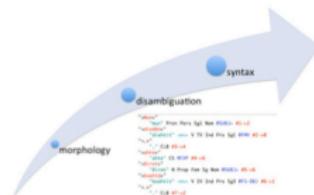
Corpus



Lexicon



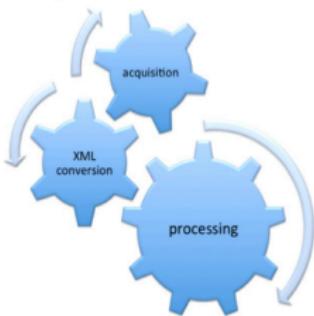
Analysis



The parallel corpus will be available online for search via appropriate graphical user interfaces such as *KORP*.

The screenshot shows the KORP search interface. At the top, it says "Dependency test corpus selected — 10,305 tokens". Below that is a search bar with "Simple", "Extended", and "Advanced" tabs, currently set to "Simple". The search term "vättis" is entered, and there are checkboxes for "Search also as" and options for "initial part", "final part", and "case-insensitive". Below the search bar are buttons for "KWIC: hits per page: 25", "sort within corpora: not sorted", and "Statistics: compile based on: word". The main area has tabs for "KWIC", "Statistics", and "Word picture", with "KWIC" selected. It displays "Results: 7" and a link "Show context". A modal window titled "Dependency Tree" is open, showing the dependency relations for the sentence "Dat ii leat gal nu vättis.". The tree diagram shows "Munne le" as the root node, with children "Pron" (SUB), "ii" (MV), "leat" (V), "gal" (ADV), "nu" (ADV), and "vättis" (CLB). Arrows indicate dependencies between these words. The background shows the full sentence "Dat ii leat gal nu vättis." repeated twice, with the last word "vättis" highlighted.

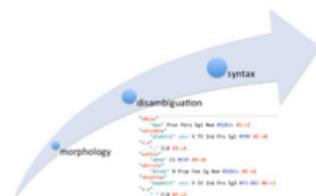
Corpus



Lexicon



Analysis



Search

The screenshot shows the KORP interface with the following details:

- Header:** Dependency test corpus selected - 62,000 tokens
- Search:** Search for `WDT`, search as also, initial part, final part and case-insensitive
- Statistics:** KWIC, hits per page, 25; sort entries, corpora, not sorted; Statistics; compile based on word
- Panel:** KWIC, Statistics, Word pictures
- Results:** 7 results found.
- Show context:** Det kan gal bli vitt.
- Dependency Tree:**
 - Root node: Det
 - Dependents: kan, gal, bli, vitt
 - Relationships:
 - kan: Dependency relation, agent, ADV
 - gal: Dependency relation, modif, NOUN
 - bli: Dependency relation, pred, VERB
 - vitt: Dependency relation, object, NOUN
- Morphology:** Det, kan, gal, bli, vitt



For Computer-Assisted Translation (CAT), both a parallel corpus as Translation Memory (TM) and a glossary with the FAD lexical material have been built.

Autoshumato ITE :: st_åm

The screenshot shows the Autoshumato ITE interface with two main panes. The left pane displays a document titled 'Editor - st_åm.odt' containing Sametinget's annual report for 2011. The right pane contains three tabs: 'Fuzzy Matches', 'Glossary', and 'Dictionary'. The 'Fuzzy Matches' tab lists several matches from the TM, such as 'Lov og forskrift om offentlige anskaffelser ble endret i 2001.' and 'Horing – Sametingsvalget – forskrift om valg til Sametinget'. The 'Glossary' tab lists terms like 'forskrift = láhkáásahus', 'valg = válga', and 'til = dassá'. The 'Dictionary' tab is currently empty.

Editor - st_åm.odt

Sametingets årsmeddelelse omratter Sametingets politiske og forvaltningsmessige aktivitet i 2011.

Årsmeldingen rapporterer om oppfølgingen av hovedmålene og delmålene for de enkelte fagområdene og bruken av virkemidlene i budsjettet for 2011.

Sametingets regnskap for 2011 er vedlagt årsmeldingen.

Sametingsvalget
Sámediggeválggat

Forskrift om valg til Sametinget ble endret i 2011.
<segment 0013>

Endringen innebærer at Sametingets valgmannstall også utarbeides i de årene det er kommune- og fylkestingsvalg.

Dette valgmannstallet bestemmer hvilke kommuner som skal avholde valgting ved neste sametingsvalg, hvilke kommuner som kun har forhåndsstemmegivning og mandatfordelingen på valgkretsene.

Forskriftsendringene er en oppfølging av plenumsvedtak fra 2007 om valgordningen til sametingsvalget.

14 162 personer var i 2011 innmeldt i Sametingets valgmannstall.

Det er ni nye kommuner som får valgting, og ved sametingsvalget i 2013 vil i alt 56 kommuner ha valgting.

Fuzzy Matches

- 1) Lov og forskrift om offentlige anskaffelser ble endret i 2001.
Almmolaš oastimaid láhka ja láhkáásahus rievdaduvvui 2001:s.
<44/44/60% /Users/ttr000/Documents/tm/nob2sme-2012-10-21/nob2sme-2012-10-21-000679.tmx >
- 2) Horing – Sametingsvalget – forskrift om valg til Sametinget
Sámediggeválgga – láhkáásahus sámediggeválgga birra – gulaskuddan
<37/37/44% /Users/ttr000/Documents/tm/nob2sme-2012-10-21/nob2sme-2012-10-21-000894.tmx >
- 3) Loven ble endret i 2006.
Láhka – rievdaduvvui jagi 2006.
<37/37/44% /Users/ttr000/Documents/tm/nob2sme-2012-10-21/nob2sme-2012-10-21-000032.tmx >

Glossary

forskrift = láhkáásahus
valg = válga
til = dassá
Sametinget = Sámediggi

Dictionary

--

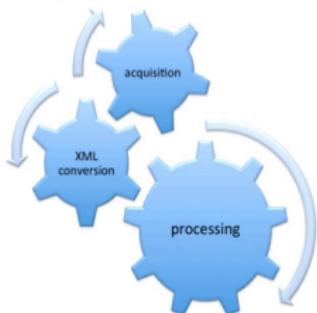
Machine Translation | Multiple Translations | Notes | Comments

Project autosaved on 7:34 PM

27/4118 (18/3044, 4119) 51/0



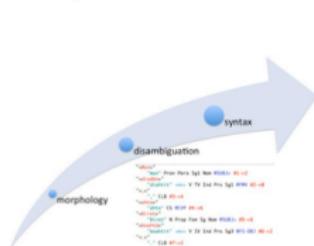
Corpus



Lexicon



Analysis



Search

A screenshot of the KORP search interface. The search bar contains 'Search for: Dat'. Below it is a table with columns 'KMC', 'Statistics', and 'Word picture'. The 'Word picture' section shows a dependency tree for the word 'Dat', with nodes like 'Det' and 'ADV' and arrows indicating grammatical relations.

CAT

A screenshot of the CAT (Computer Assisted Translation) interface. The search bar contains 'Søg i kataloget'. Below it is a table with columns 'Titel', 'Forfatter', 'Udgivelsesår', and 'Genre'. A detailed view of a document entry is shown, including fields like 'Titel', 'Forfatter', 'Udgivelsesår', 'Genre', 'Bemærkninger', and 'Forside'.



The acquired knowledge can be used for other Saami languages.

Bures boahtin!



Buerie båeteme!

Know-how

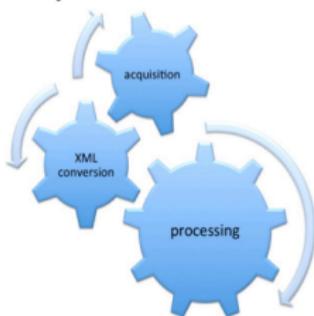
- Corpus
- Tools
- Pipelines

How?

- ...
- ...
- ...

A good teamwork is essential to the final result.

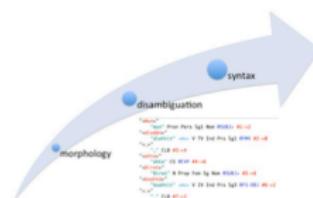
Corpus



Lexicon



Analysis



Search

The Search section displays the KORP interface with the following details:

- Title**: Dependency test corpus selected - 10,000 entries
- Search for**: **VEER**
- Results**: 7
- Dependency Tree** (example):


```

      Dat is heel gaaf dat is heel.
      ┌─────────┐ ┌─────────┐
      |         | |         |
      └────────┘ └────────┘
      Dat   is   heel  gaaf  dat   is   heel.
      ┌─────────┐ ┌─────────┐ ┌─────────┐ ┌─────────┐ ┌─────────┐ ┌─────────┐
      |         | |         | |         | |         | |         |
      └────────┘ └────────┘ └────────┘ └────────┘ └────────┘ └────────┘
      Dat  is  heel  gaaf  dat  is  heel.
      ┌─────────┐ ┌─────────┐ ┌─────────┐ ┌─────────┐ ┌─────────┐ ┌─────────┐
      |         | |         | |         | |         | |         |
      └────────┘ └────────┘ └────────┘ └────────┘ └────────┘ └────────┘
      Dat  is  heel  gaaf  dat  is  heel.
      ┌─────────┐ ┌─────────┐ ┌─────────┐ ┌─────────┐ ┌─────────┐ ┌─────────┐
      |         | |         | |         | |         | |         |
      └────────┘ └────────┘ └────────┘ └────────┘ └────────┘ └────────┘
      
```

CAT

The CAT section displays the following components:

- Glossary** (example):

verb	verb
Substantiviserende verb (verb)	Substantiviserende verb (verb)
- Search history** (example):
 - 1. Nering - Semantipunkt - berättar om vad som hänt!
 - 2. Nering - Semantipunkt - berättar om vad som hänt!
 - 3. Nering - Semantipunkt - berättar om vad som hänt!
 - 4. Nering - Semantipunkt - berättar om vad som hänt!
 - 5. Nering - Semantipunkt - berättar om vad som hänt!
 - 6. Nering - Semantipunkt - berättar om vad som hänt!
 - 7. Nering - Semantipunkt - berättar om vad som hänt!
 - 8. Nering - Semantipunkt - berättar om vad som hänt!
 - 9. Nering - Semantipunkt - berättar om vad som hänt!
 - 10. Nering - Semantipunkt - berättar om vad som hänt!
 - 11. Nering - Semantipunkt - berättar om vad som hänt!
 - 12. Nering - Semantipunkt - berättar om vad som hänt!
 - 13. Nering - Semantipunkt - berättar om vad som hänt!
 - 14. Nering - Semantipunkt - berättar om vad som hänt!
 - 15. Nering - Semantipunkt - berättar om vad som hänt!
 - 16. Nering - Semantipunkt - berättar om vad som hänt!
 - 17. Nering - Semantipunkt - berättar om vad som hänt!
 - 18. Nering - Semantipunkt - berättar om vad som hänt!
 - 19. Nering - Semantipunkt - berättar om vad som hänt!
 - 20. Nering - Semantipunkt - berättar om vad som hänt!
 - 21. Nering - Semantipunkt - berättar om vad som hänt!
 - 22. Nering - Semantipunkt - berättar om vad som hänt!
 - 23. Nering - Semantipunkt - berättar om vad som hänt!
 - 24. Nering - Semantipunkt - berättar om vad som hänt!
 - 25. Nering - Semantipunkt - berättar om vad som hänt!
 - 26. Nering - Semantipunkt - berättar om vad som hänt!
 - 27. Nering - Semantipunkt - berättar om vad som hänt!
 - 28. Nering - Semantipunkt - berättar om vad som hänt!
 - 29. Nering - Semantipunkt - berättar om vad som hänt!
 - 30. Nering - Semantipunkt - berättar om vad som hänt!
 - 31. Nering - Semantipunkt - berättar om vad som hänt!
 - 32. Nering - Semantipunkt - berättar om vad som hänt!
 - 33. Nering - Semantipunkt - berättar om vad som hänt!
 - 34. Nering - Semantipunkt - berättar om vad som hänt!
 - 35. Nering - Semantipunkt - berättar om vad som hänt!
 - 36. Nering - Semantipunkt - berättar om vad som hänt!
 - 37. Nering - Semantipunkt - berättar om vad som hänt!
 - 38. Nering - Semantipunkt - berättar om vad som hänt!
 - 39. Nering - Semantipunkt - berättar om vad som hänt!
 - 40. Nering - Semantipunkt - berättar om vad som hänt!
 - 41. Nering - Semantipunkt - berättar om vad som hänt!
 - 42. Nering - Semantipunkt - berättar om vad som hänt!
 - 43. Nering - Semantipunkt - berättar om vad som hänt!
 - 44. Nering - Semantipunkt - berättar om vad som hänt!
 - 45. Nering - Semantipunkt - berättar om vad som hänt!
 - 46. Nering - Semantipunkt - berättar om vad som hänt!
 - 47. Nering - Semantipunkt - berättar om vad som hänt!
 - 48. Nering - Semantipunkt - berättar om vad som hänt!
 - 49. Nering - Semantipunkt - berättar om vad som hänt!
 - 50. Nering - Semantipunkt - berättar om vad som hänt!
 - 51. Nering - Semantipunkt - berättar om vad som hänt!
 - 52. Nering - Semantipunkt - berättar om vad som hänt!
 - 53. Nering - Semantipunkt - berättar om vad som hänt!
 - 54. Nering - Semantipunkt - berättar om vad som hänt!
 - 55. Nering - Semantipunkt - berättar om vad som hänt!
 - 56. Nering - Semantipunkt - berättar om vad som hänt!
 - 57. Nering - Semantipunkt - berättar om vad som hänt!
 - 58. Nering - Semantipunkt - berättar om vad som hänt!
 - 59. Nering - Semantipunkt - berättar om vad som hänt!
 - 60. Nering - Semantipunkt - berättar om vad som hänt!
 - 61. Nering - Semantipunkt - berättar om vad som hänt!
 - 62. Nering - Semantipunkt - berättar om vad som hänt!
 - 63. Nering - Semantipunkt - berättar om vad som hänt!
 - 64. Nering - Semantipunkt - berättar om vad som hänt!
 - 65. Nering - Semantipunkt - berättar om vad som hänt!
 - 66. Nering - Semantipunkt - berättar om vad som hänt!
 - 67. Nering - Semantipunkt - berättar om vad som hänt!
 - 68. Nering - Semantipunkt - berättar om vad som hänt!
 - 69. Nering - Semantipunkt - berättar om vad som hänt!
 - 70. Nering - Semantipunkt - berättar om vad som hänt!
 - 71. Nering - Semantipunkt - berättar om vad som hänt!
 - 72. Nering - Semantipunkt - berättar om vad som hänt!
 - 73. Nering - Semantipunkt - berättar om vad som hänt!
 - 74. Nering - Semantipunkt - berättar om vad som hänt!
 - 75. Nering - Semantipunkt - berättar om vad som hänt!
 - 76. Nering - Semantipunkt - berättar om vad som hänt!
 - 77. Nering - Semantipunkt - berättar om vad som hänt!
 - 78. Nering - Semantipunkt - berättar om vad som hänt!
 - 79. Nering - Semantipunkt - berättar om vad som hänt!
 - 80. Nering - Semantipunkt - berättar om vad som hänt!
 - 81. Nering - Semantipunkt - berättar om vad som hänt!
 - 82. Nering - Semantipunkt - berättar om vad som hänt!
 - 83. Nering - Semantipunkt - berättar om vad som hänt!
 - 84. Nering - Semantipunkt - berättar om vad som hänt!
 - 85. Nering - Semantipunkt - berättar om vad som hänt!
 - 86. Nering - Semantipunkt - berättar om vad som hänt!
 - 87. Nering - Semantipunkt - berättar om vad som hänt!
 - 88. Nering - Semantipunkt - berättar om vad som hänt!
 - 89. Nering - Semantipunkt - berättar om vad som hänt!
 - 90. Nering - Semantipunkt - berättar om vad som hänt!
 - 91. Nering - Semantipunkt - berättar om vad som hänt!
 - 92. Nering - Semantipunkt - berättar om vad som hänt!
 - 93. Nering - Semantipunkt - berättar om vad som hänt!
 - 94. Nering - Semantipunkt - berättar om vad som hänt!
 - 95. Nering - Semantipunkt - berättar om vad som hänt!
 - 96. Nering - Semantipunkt - berättar om vad som hänt!
 - 97. Nering - Semantipunkt - berättar om vad som hänt!
 - 98. Nering - Semantipunkt - berättar om vad som hänt!
 - 99. Nering - Semantipunkt - berättar om vad som hänt!
 - 100. Nering - Semantipunkt - berättar om vad som hänt!

Share know-how

Bures boahtin! Buerie bæteme!

Know-how

- Corpus
- Tools
- Pipelines

How?

- ...
- ...
- ...