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

The **FAD** project

Fornyings-, **a**dministrasjons- og kirke**d**epartementet

Ođasmahttin-, hálddahus- ja girkodepartemeanta

***Motivation:* Why?**

Approach: How?

Results: What?

The **FAD** project

Fornyings-, **a**dministrasjons- og kirke**d**epartementet

Ođasmahttin-, hálddahus- ja girkodepartemeanta

***Motivation:* Why?**

***Approach:* How?**

Results: What?

The **FAD** project

Fornyings-, **a**dministrasjons- og kirke**d**epartementet

Ođasmahttin-, hálddahus- ja girkodepartemeanta

Motivation: **Why?**

Approach: **How?**

Results: **What?**

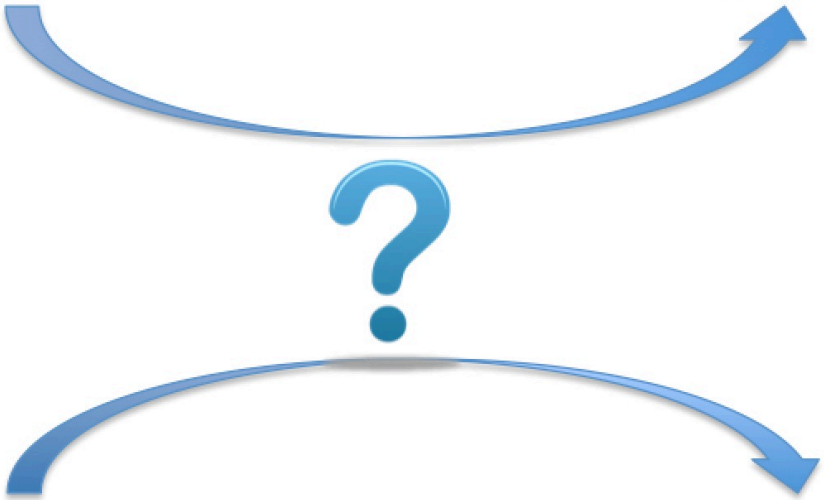
Are administrative North Saami terms exctractable from the extant Norwegian-North Saami parallel corpus?

Hensikten med ***uførepensjon*** *er...*

Jurdda lámisoajuin lea...

Jurdda buhcciidpenšuvnnain lea...

Jurdda bargonávccahisvuođaoajuin lea...

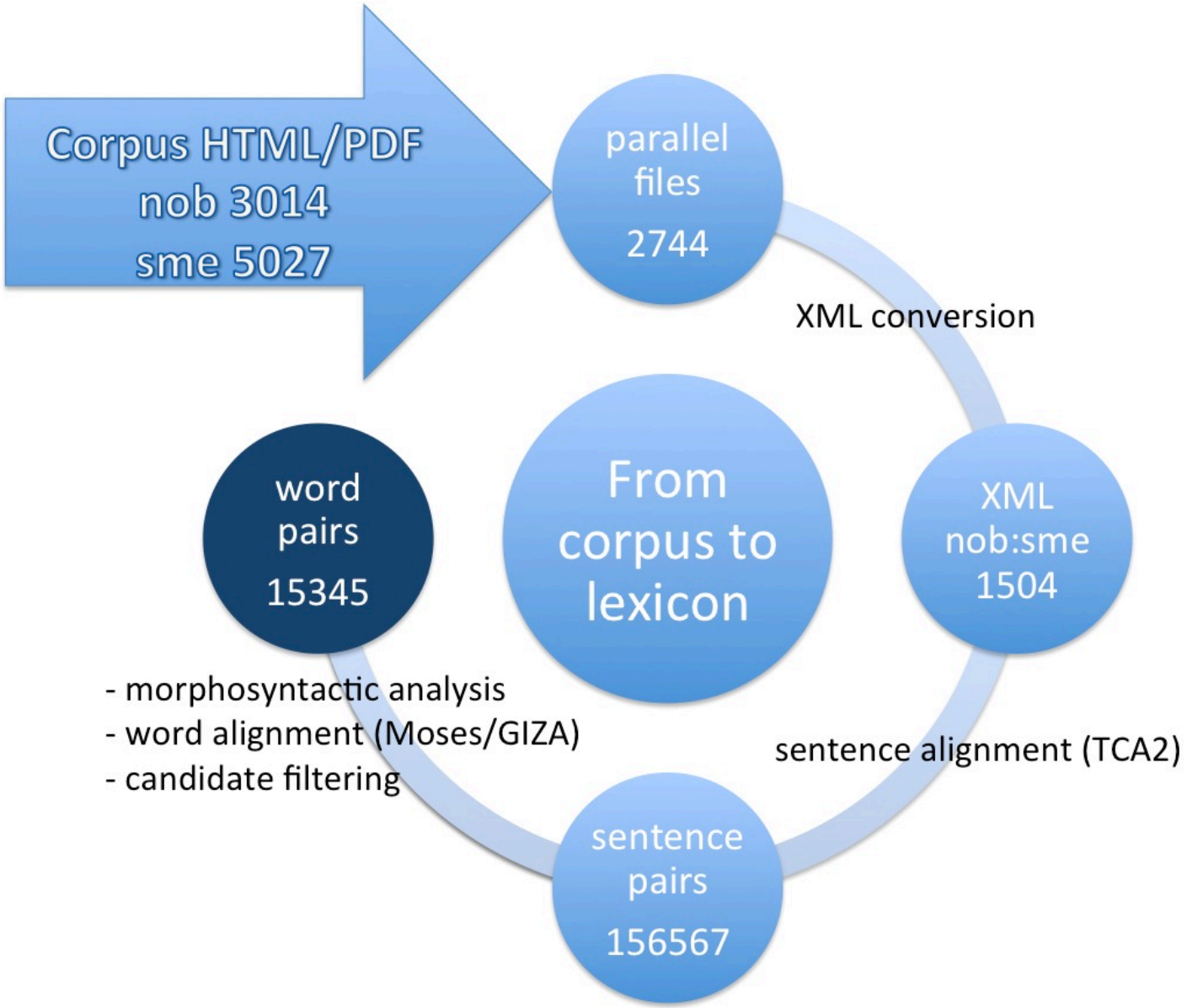


FORNYINGS-, ADMINISTRASJONS- OG KIRKEDEPARTEMENTET

ROMSSA UNIVERSITEHTA

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 0.5 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 likestillingsanalyse<subst> dásseárvoanaliisa<N>
```

```
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>
```

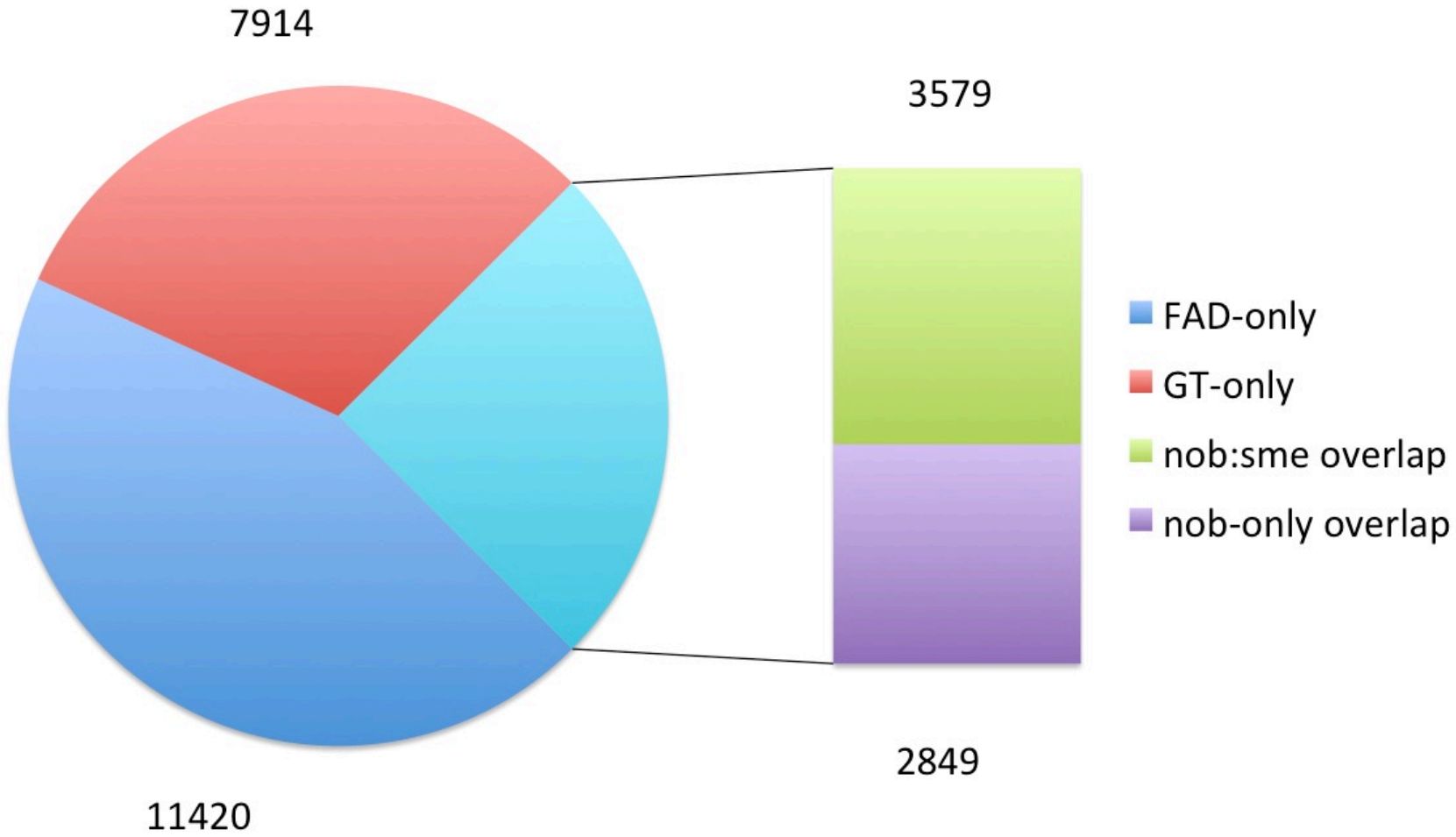
manual filtering

automatic filtering

```
1 0 -10.0 0.0 0.25 Vurderingsverktøyet<subst> árvvoštallanreaidu<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">mánnabiigá</t>
         <t pos="N" type="NomAg">mánnageahčči</t>
         <t pos="N" type="NomAg">mánnáoaidni</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">dásseárvogažaldat</t>
         <t pos="N">dásseárvoášši</t>
      </tg>
   </mg>
</e>
```

(b) FAD-only

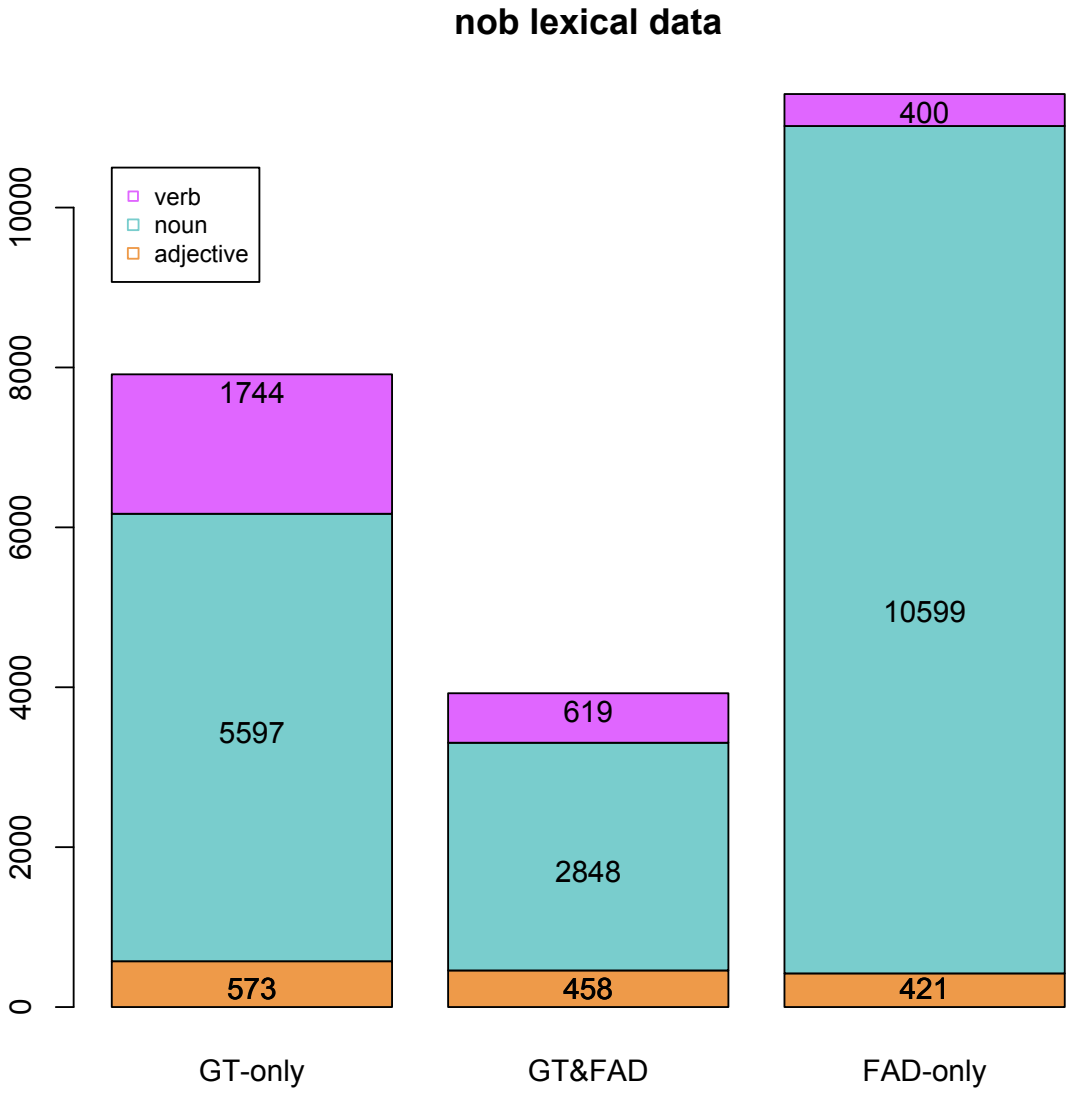
```
<e>
   <lg>
      <l pos="N">valgforbund</l>
   </lg>
   <mg>
      <tg xml:lang="sme">
         <t pos="N" src="gt;fad">válgalihttu</t>
      </tg>
   </mg>
</e>
```

(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 ovelap

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



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

How to evaluate data?

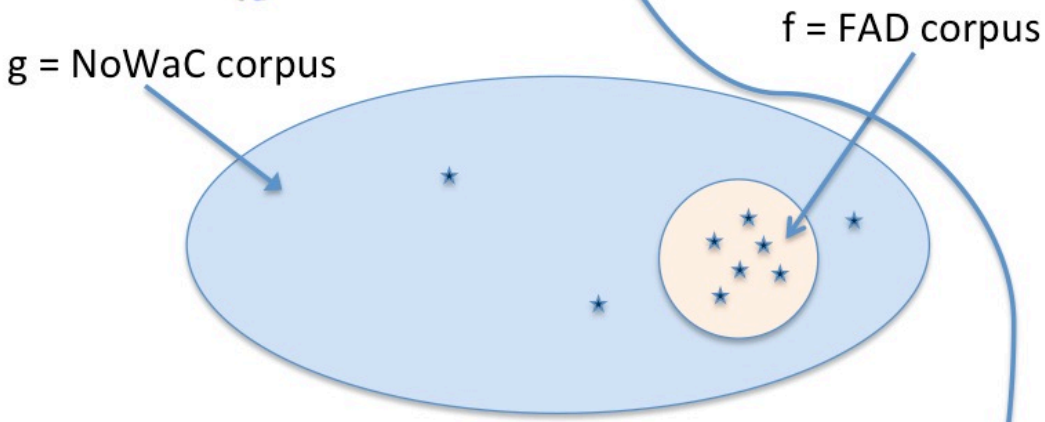
Database lookup

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

KBN
Økonomisk-
administrativ
terminologi

Term frequency

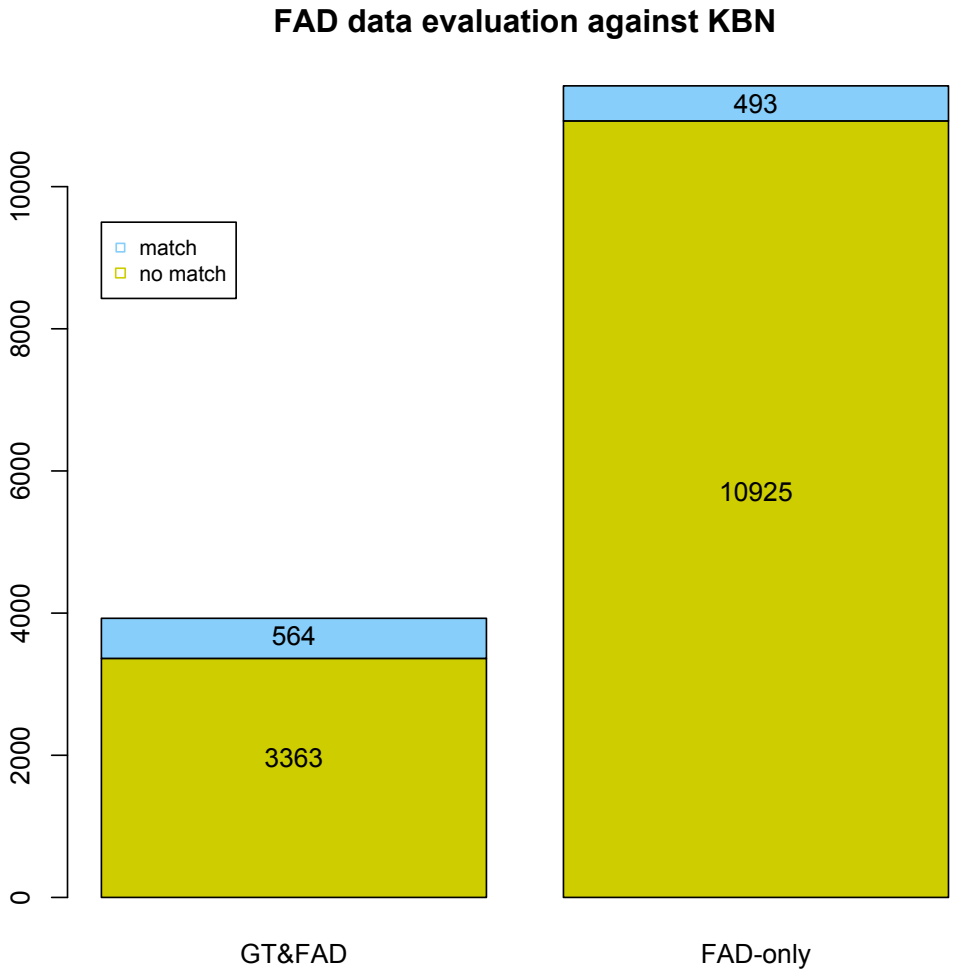
```
<e>
   <lg>
      <l pos="N" gf-ff="-0.00370519">disposisjon</l>
   </lg>
   <mg>
      <tg xml:lang="sme">
         <t pos="N" src="gt;fad" gf-ff="0.0001268">disposišuvdna</t>
      </tg>
   </mg>
   <mg>
      <tg xml:lang="sme">
         <re>jus</re>
         <t pos="N" src="fad" gf-ff="-0.00019637">háldogeavaheapmi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



gf = 0.00142683 ff = 0.00513202

gf - ff = 0.00142683 - 0.00513202 = **-0.00370519**

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

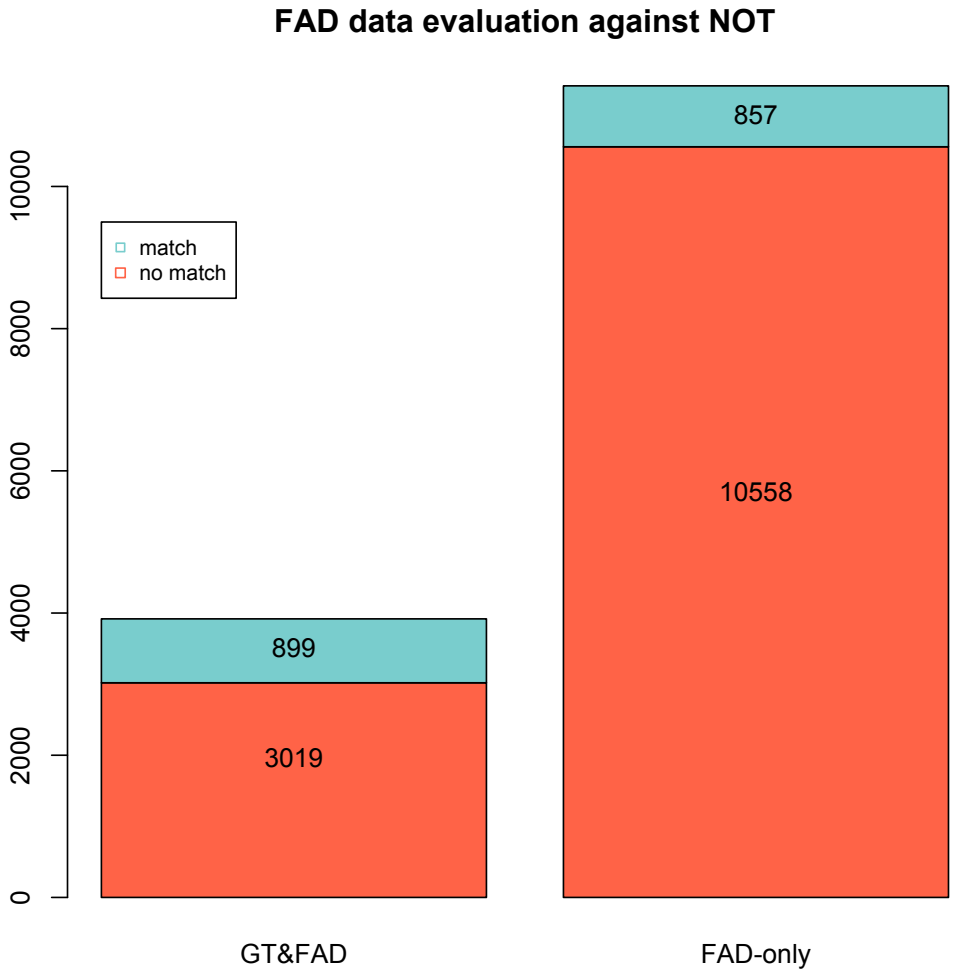


KBN termbase

- 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**.

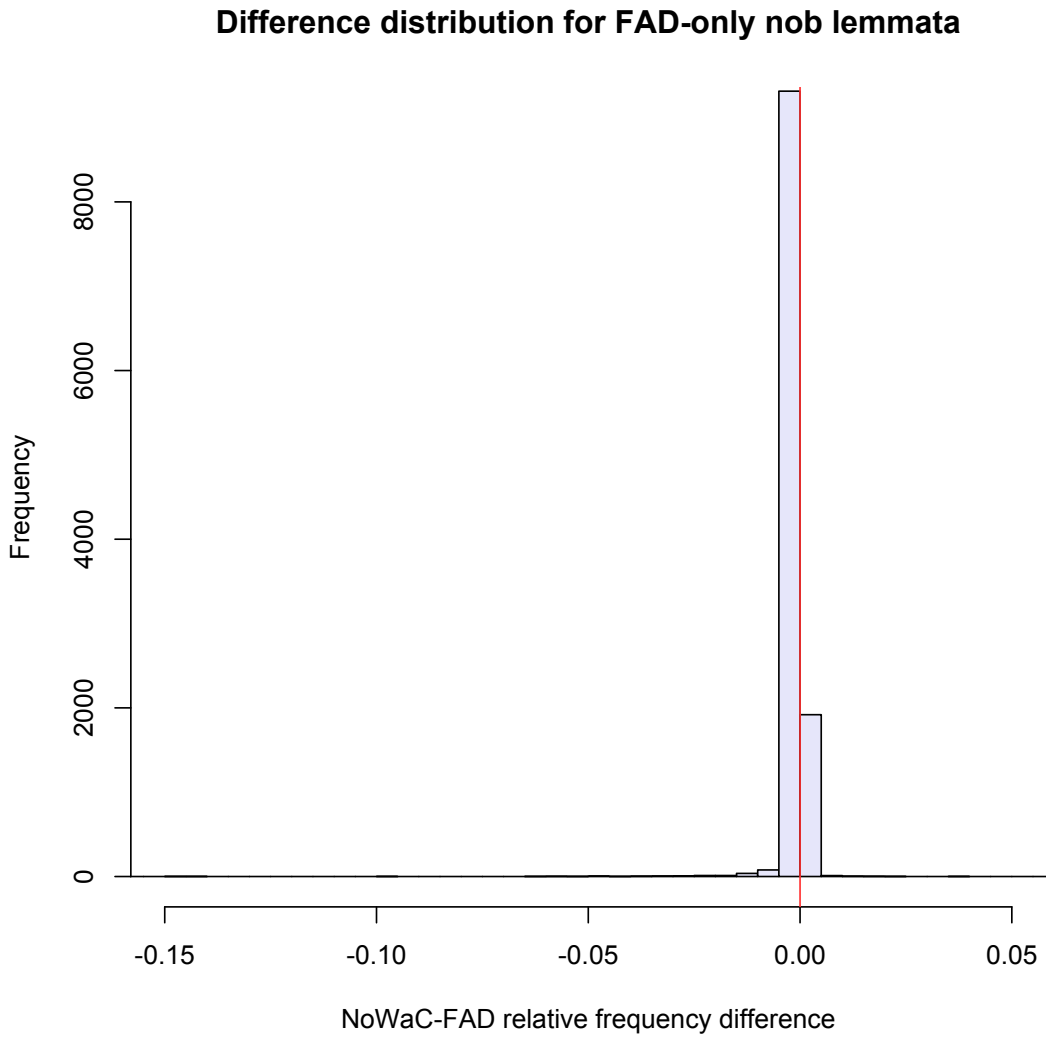


NOT termbase

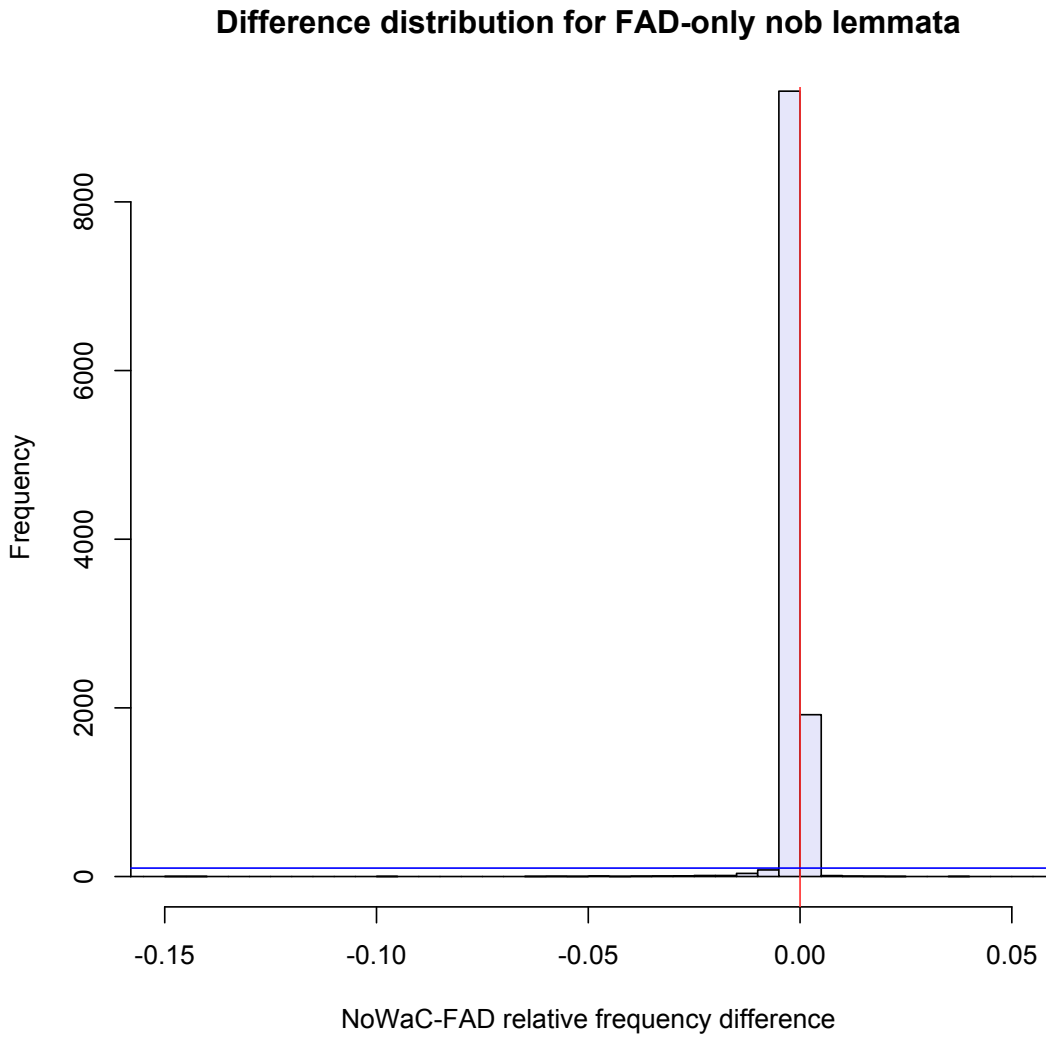
- 30525 terms
- Norsk termbank
- multilingual

  (English, Norwegian, etc.)

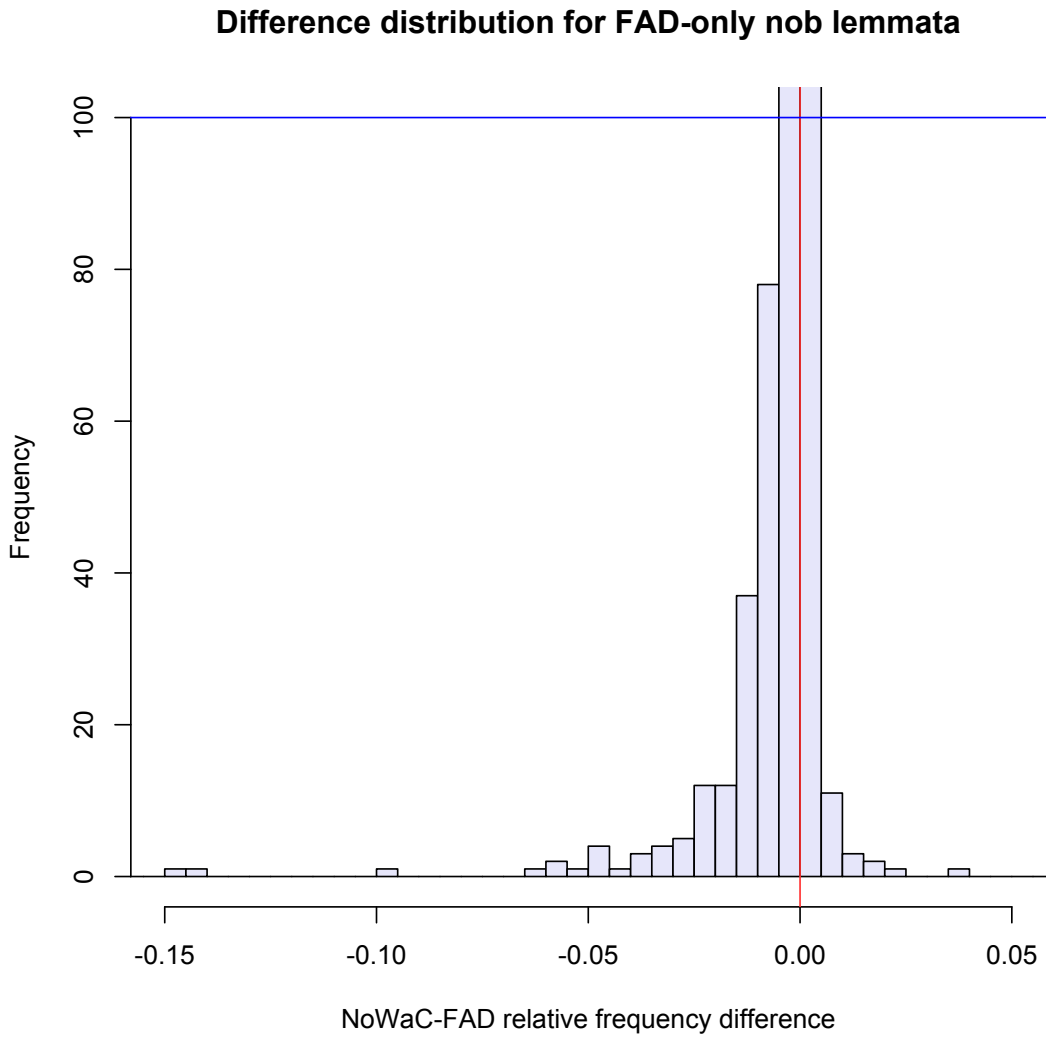
The FAD-only grouped frequency distribution shows a negative skew, more probable term candidates.



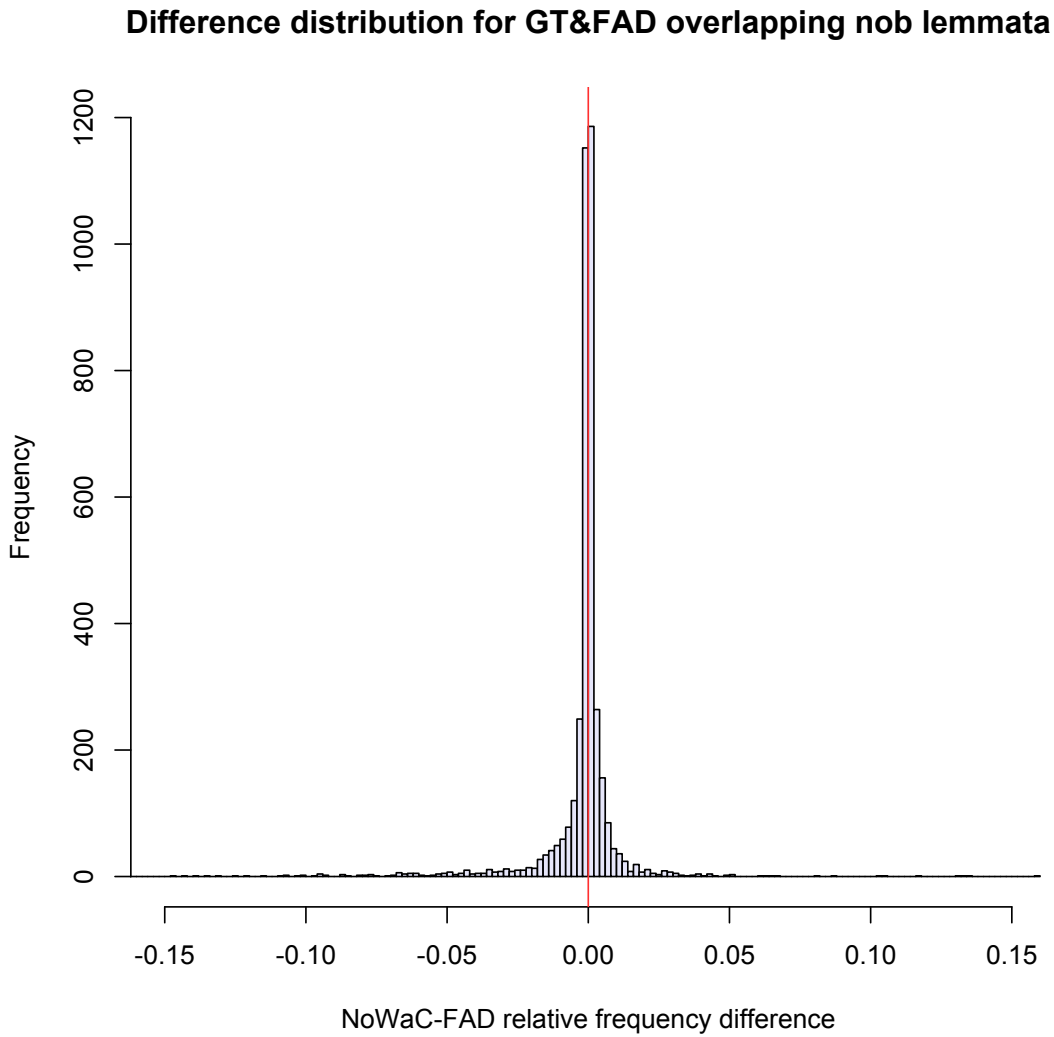
The FAD-only grouped frequency distribution shows a negative skew, more probable term candidates.



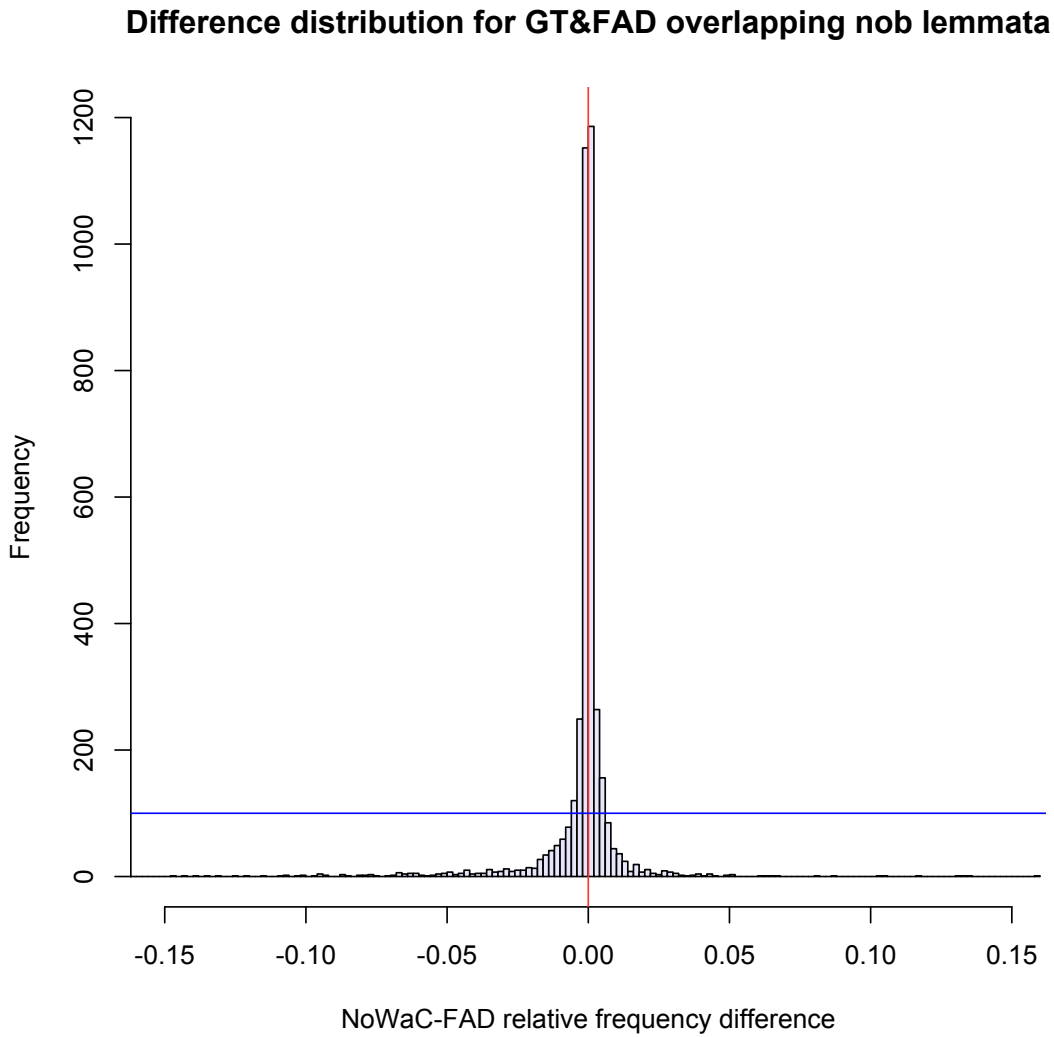
The FAD-only grouped frequency distribution shows a negative skew, more probable term candidates.



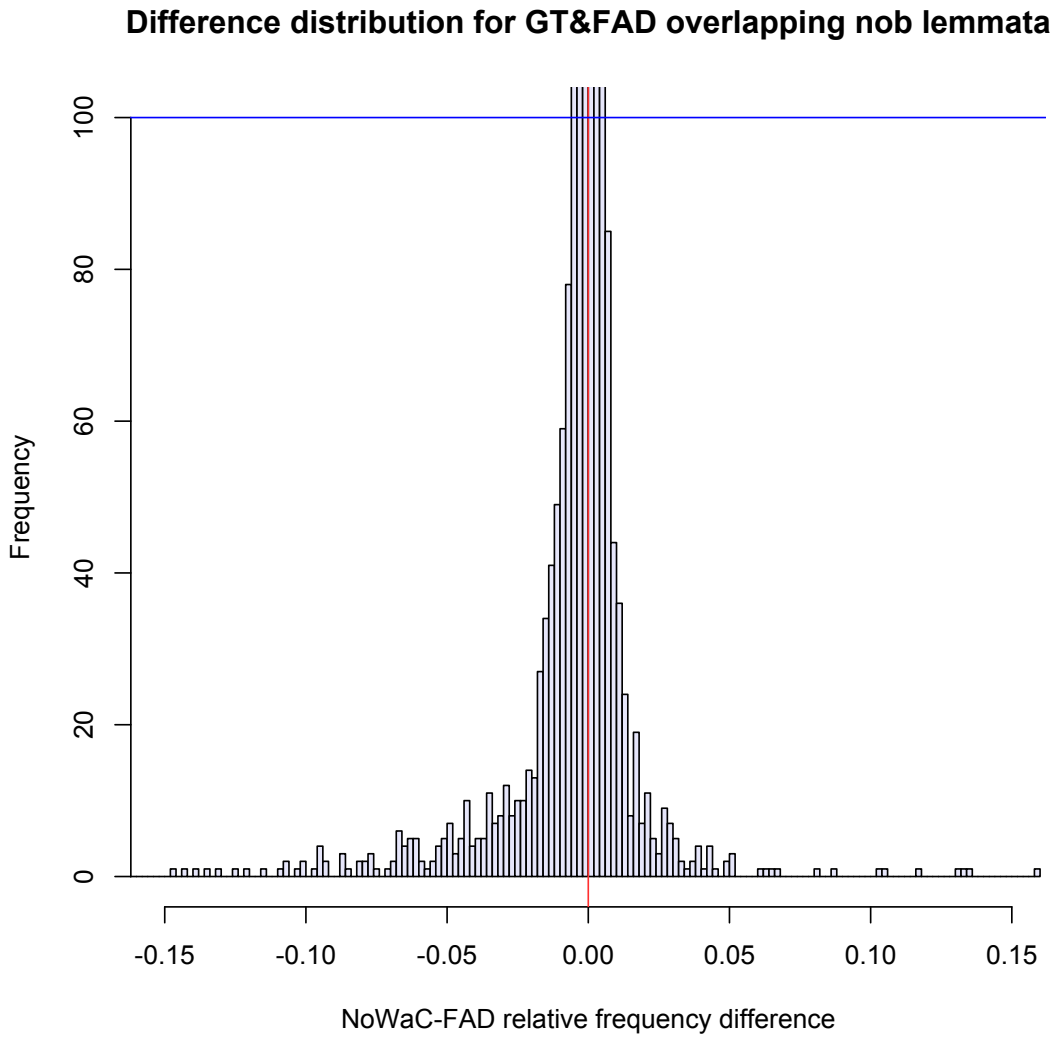
The GT&FAD grouped frequency distribution seems to be more balanced between negative and positive values.



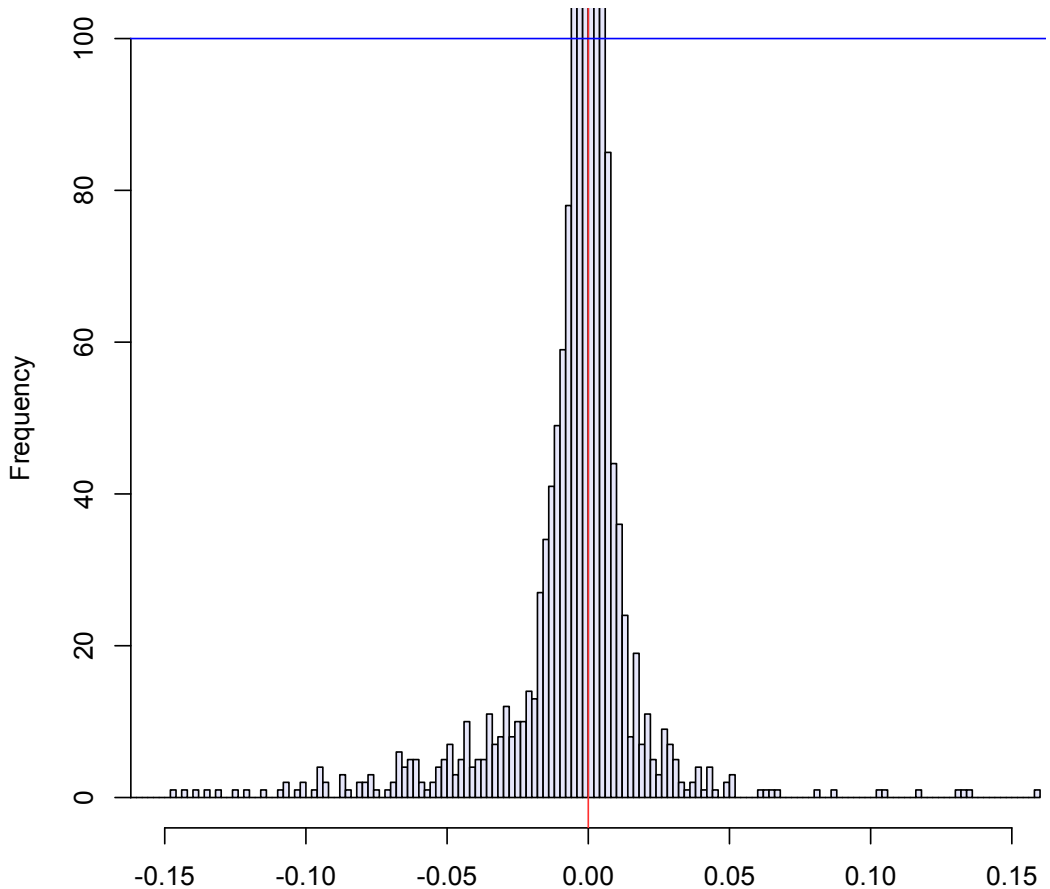
The GT&FAD grouped frequency distribution seems to be more balanced between negative and positive values.



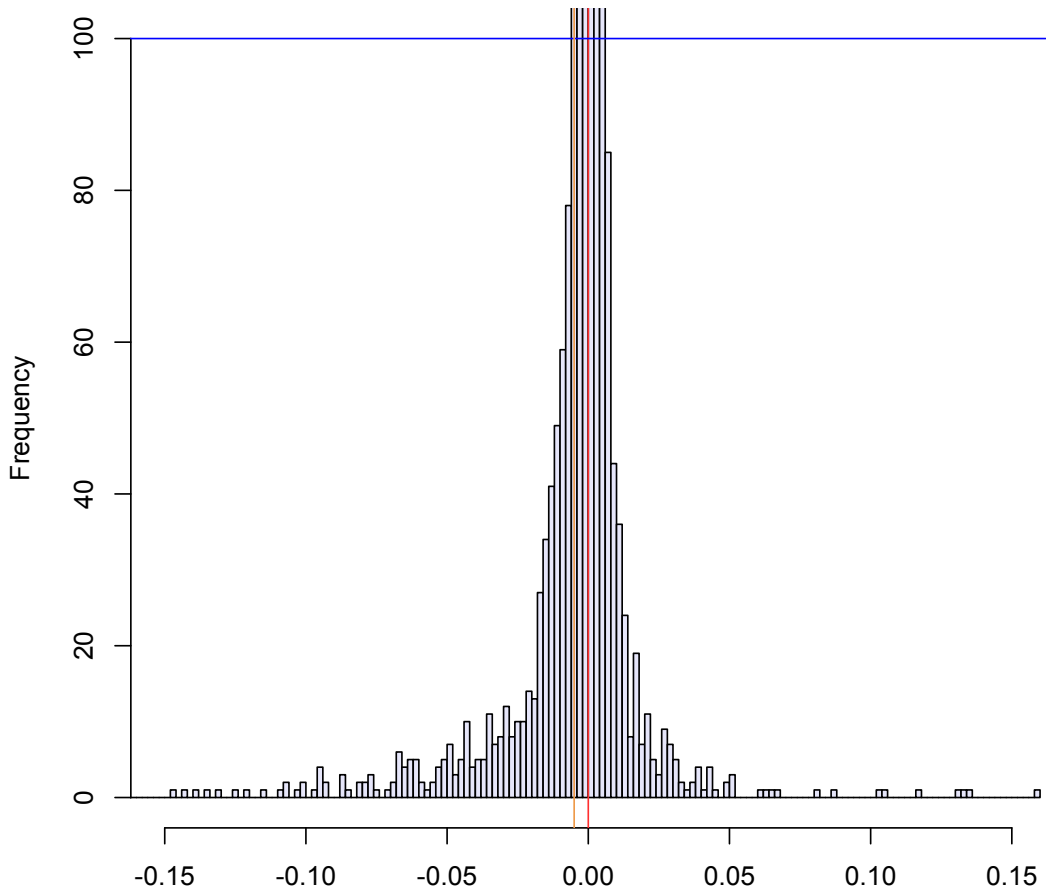
The GT&FAD grouped frequency distribution seems to be more balanced between negative and positive values.



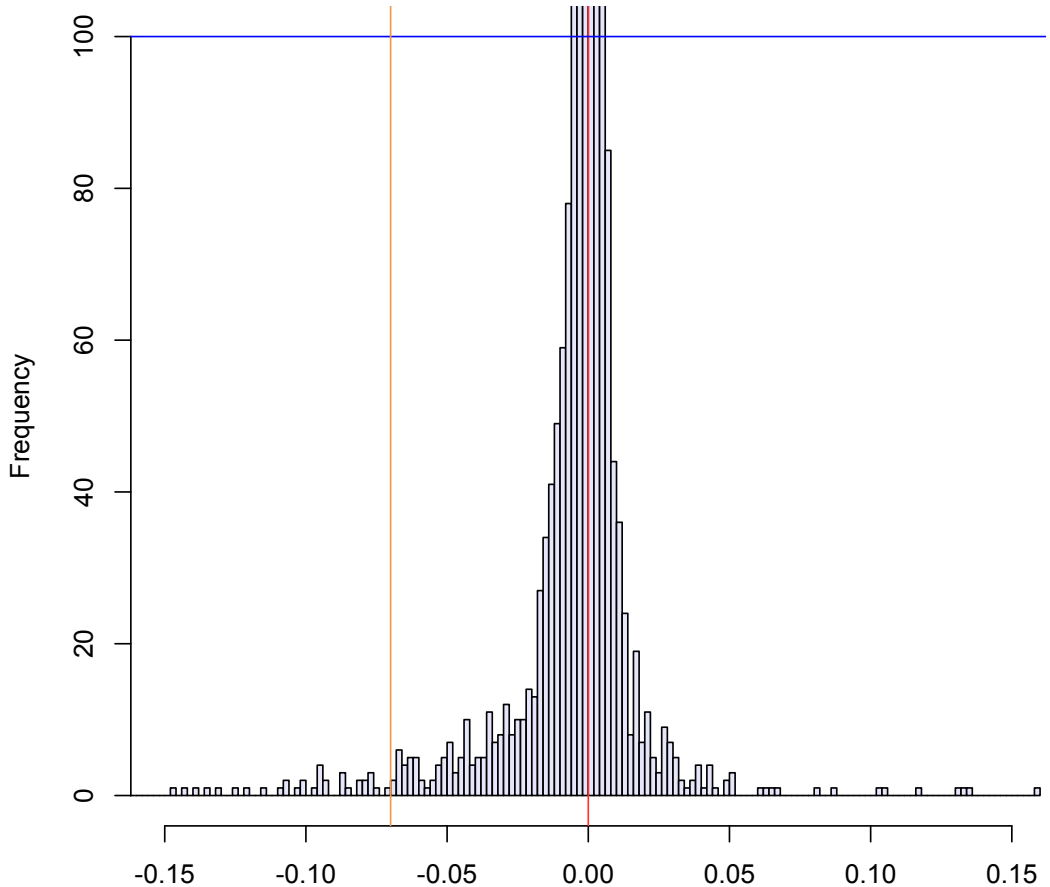
Is it possible to find a boundary between potential term candidates and ordinary words?



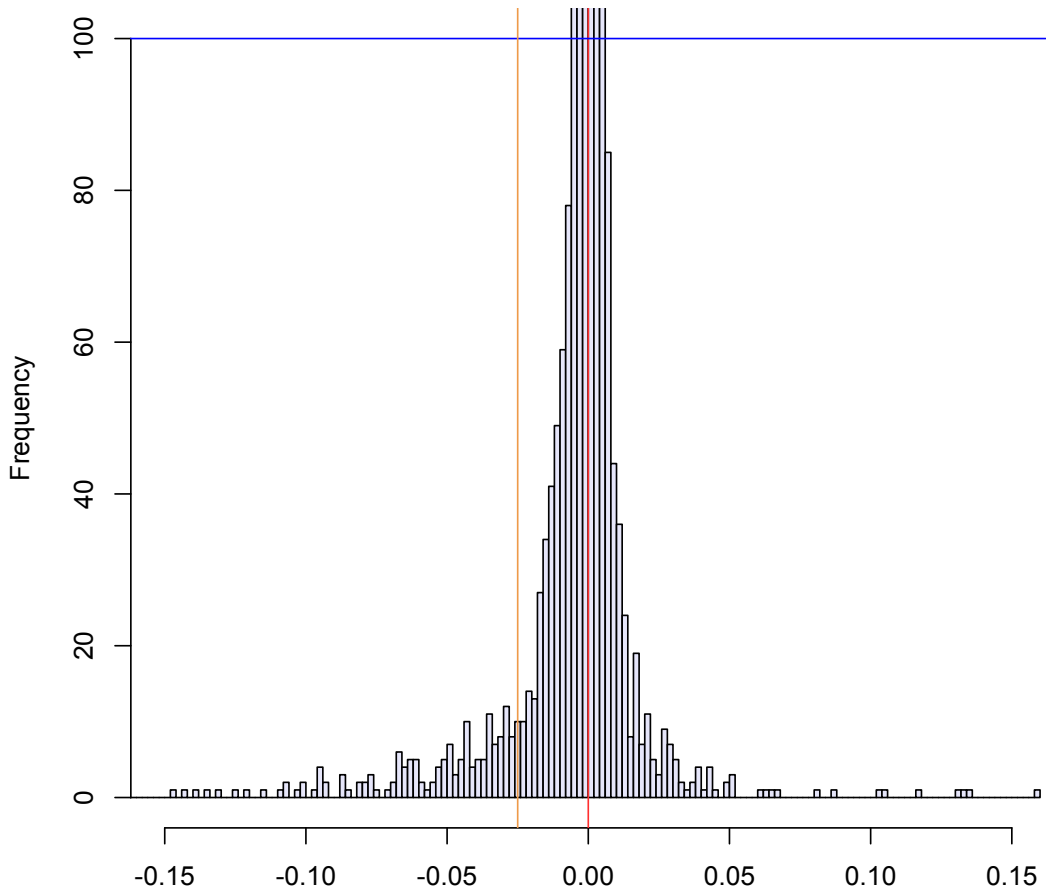
Is it possible to find a boundary between potential term candidates and ordinary words?



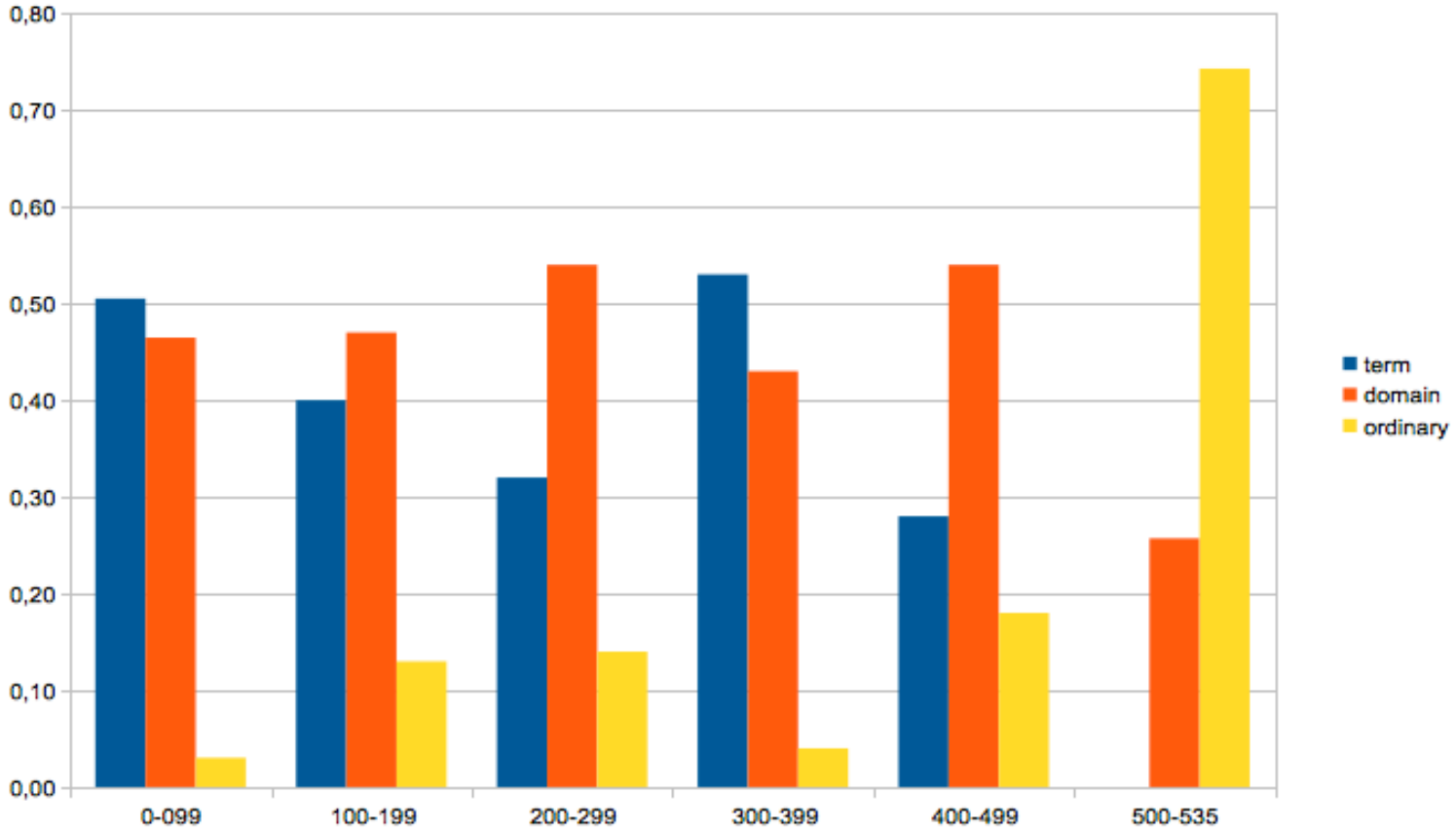
Is it possible to find a boundary between potential term candidates and ordinary words?



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.

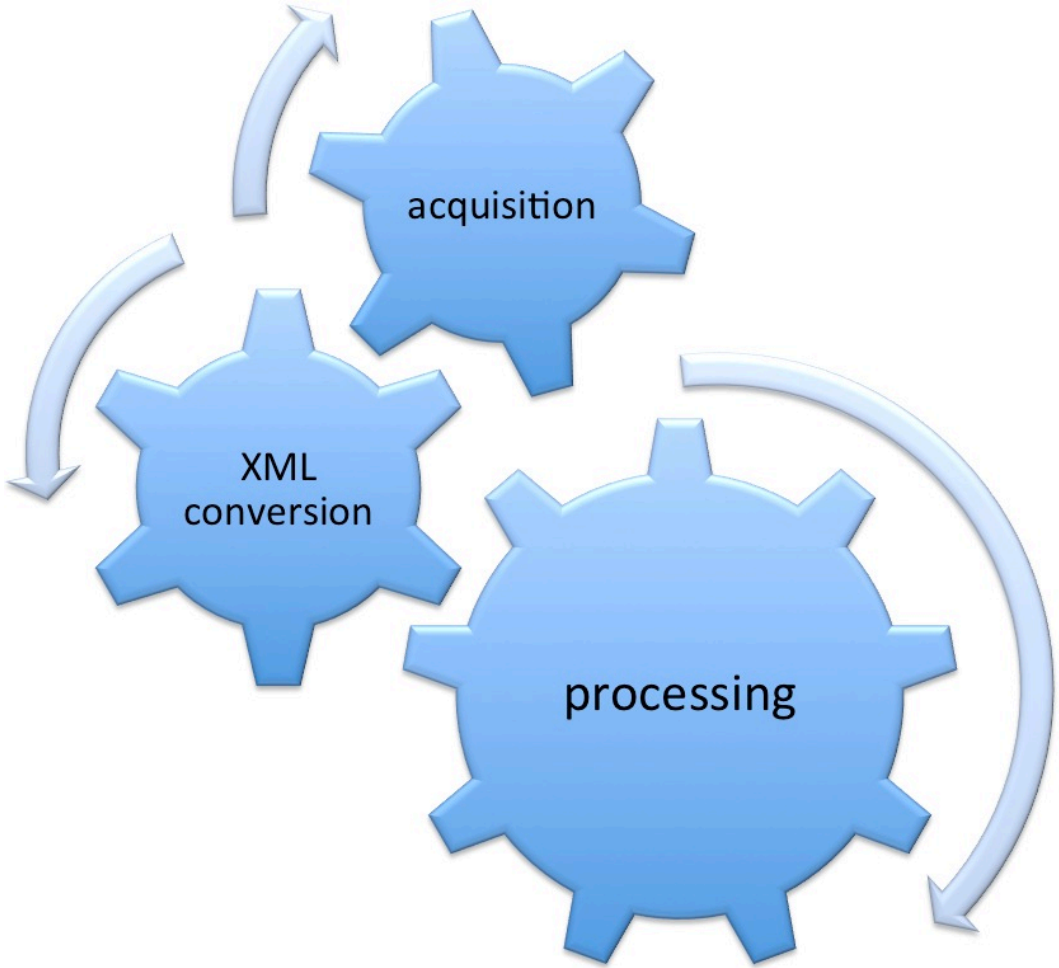


What are the results of the FAD project?

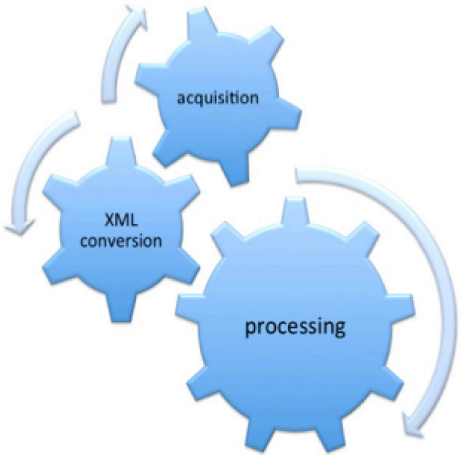




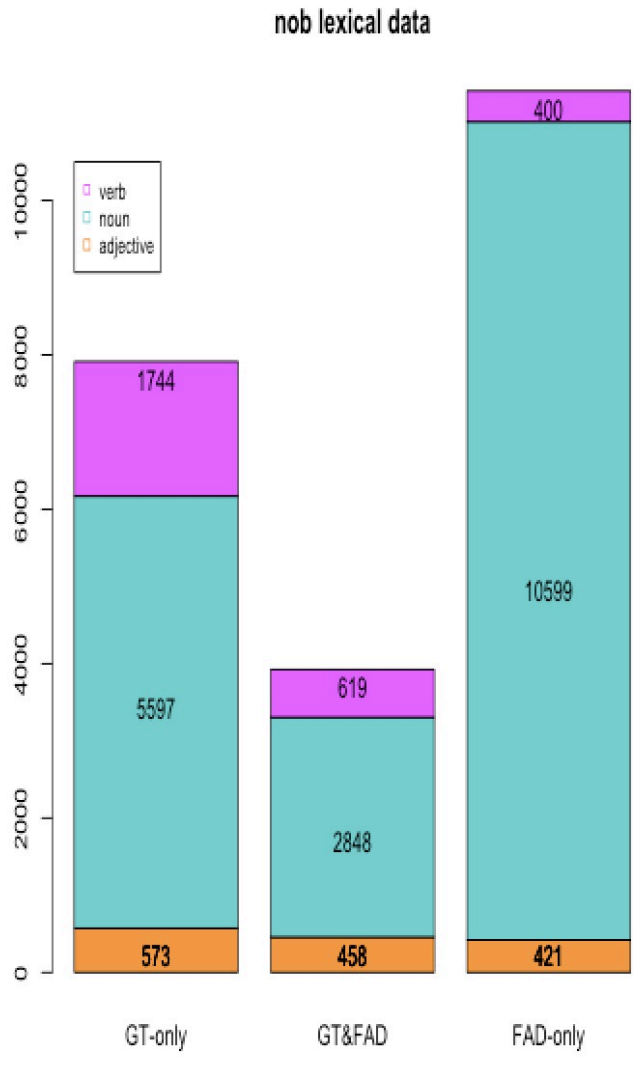

Corpus acquisition, conversion and testing routines have been greatly improved.

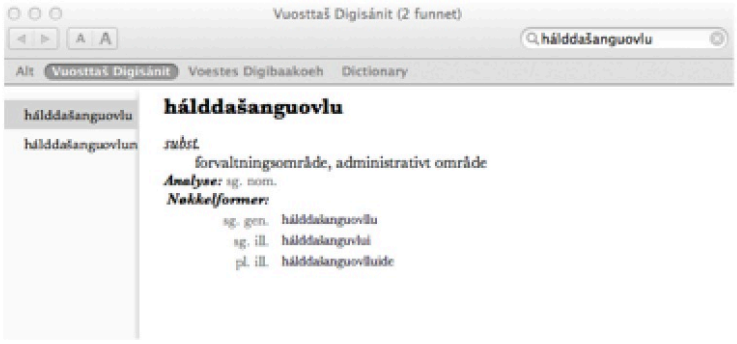


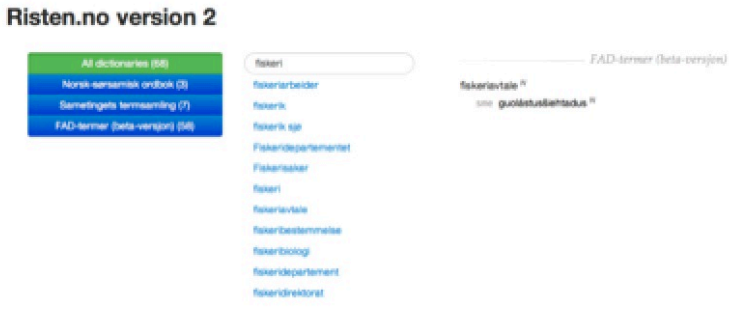
Corpus

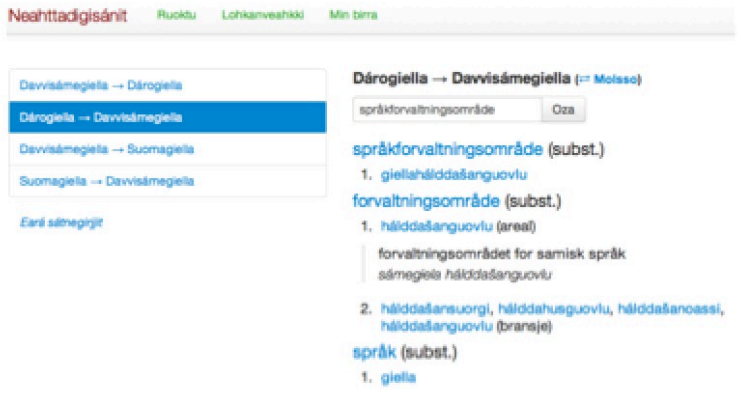


Lexica and dictionaries have been tremendously extended.

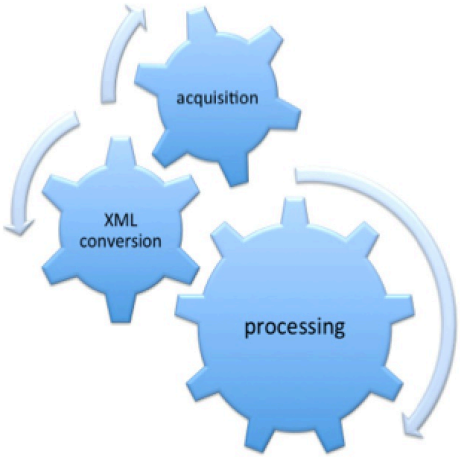






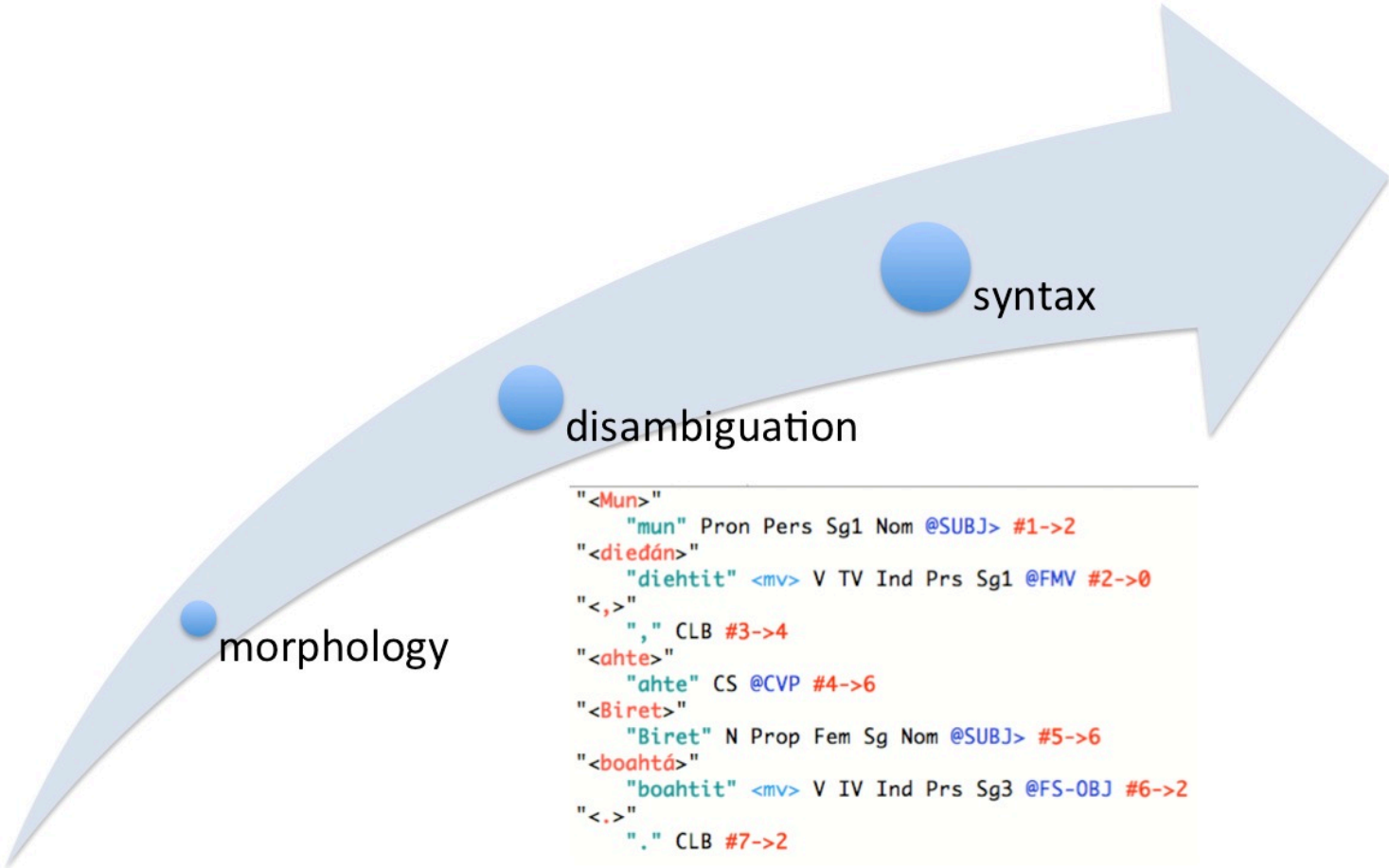


Corpus

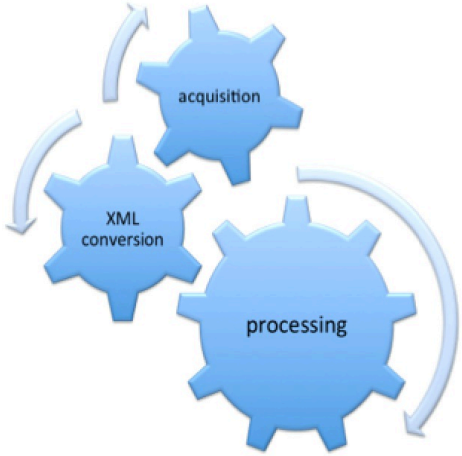


Lexicon

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

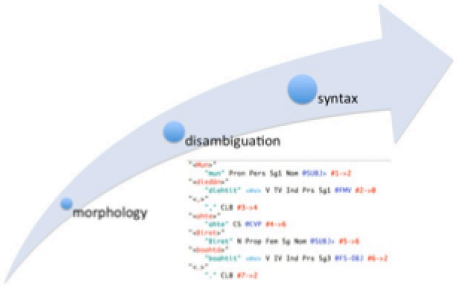


Corpus

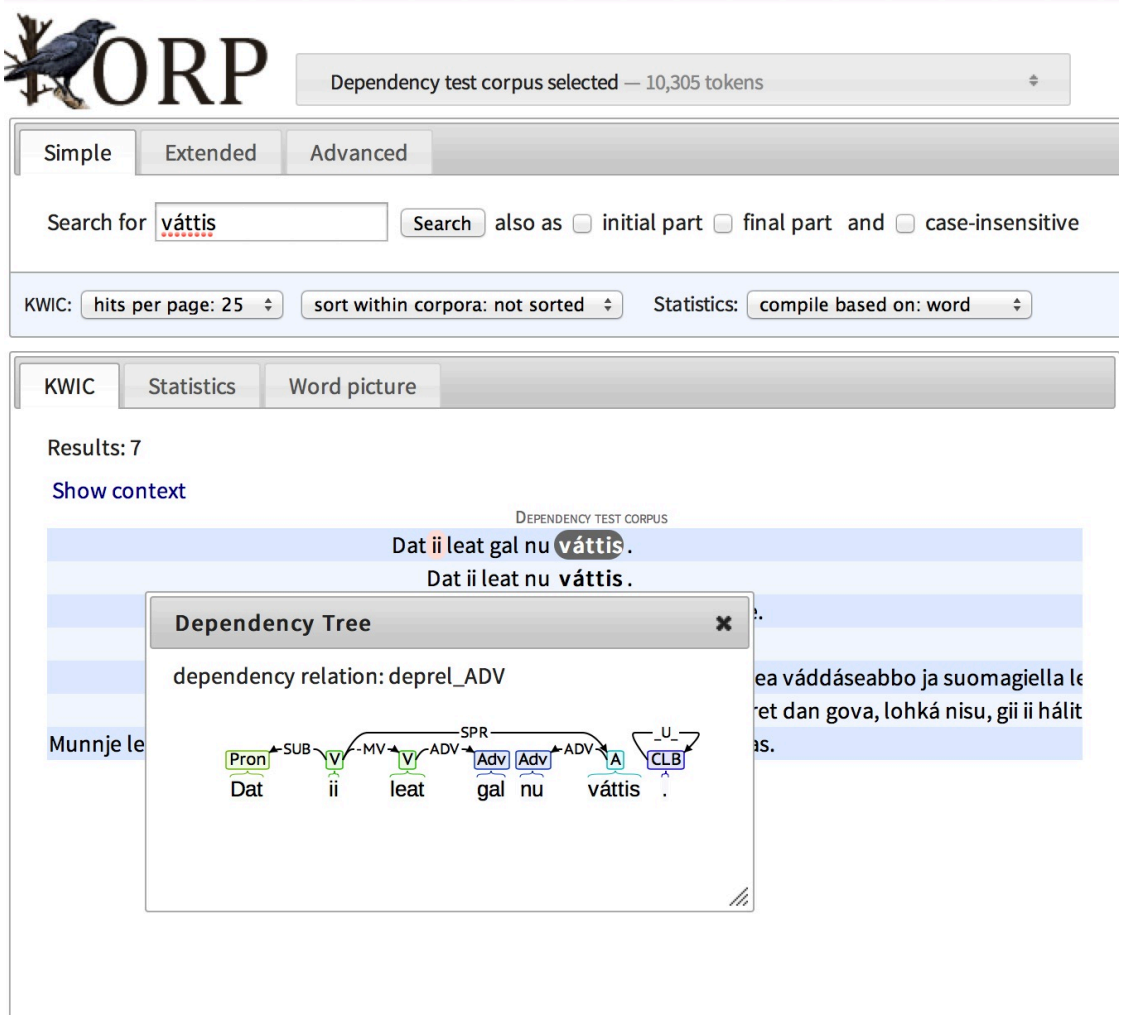


Lexicon

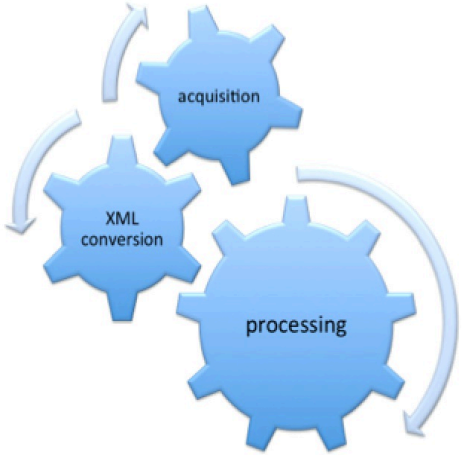
Analysis



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

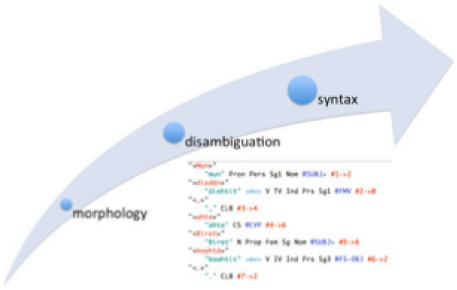


Corpus

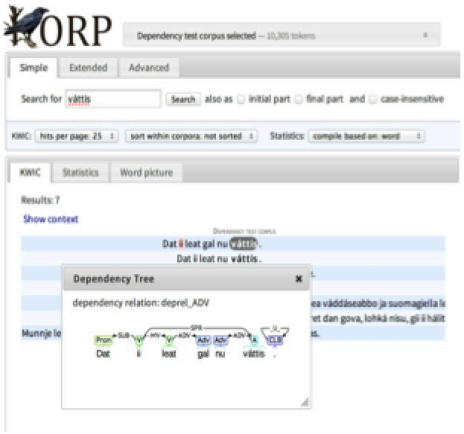


Lexicon

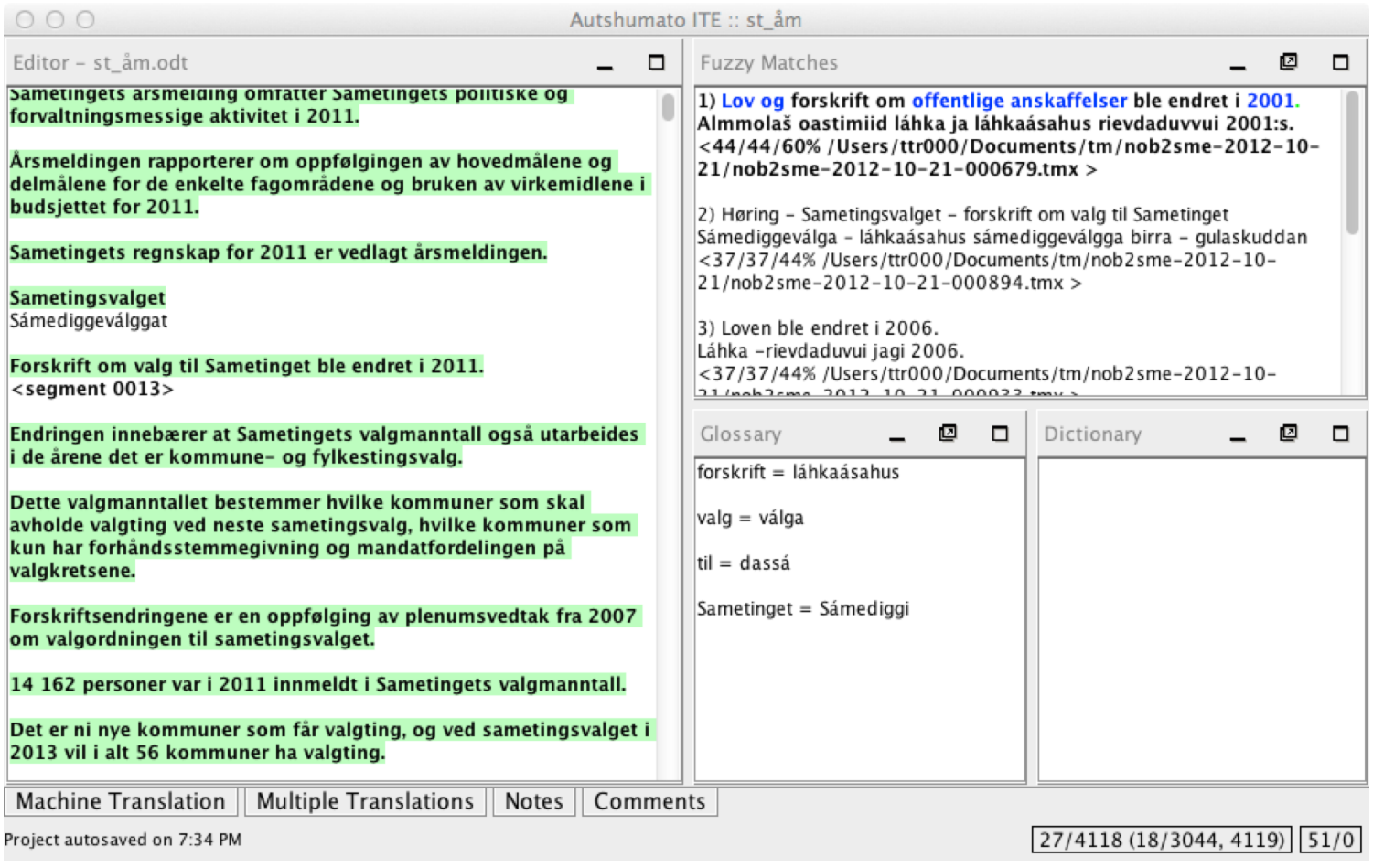
Analysis



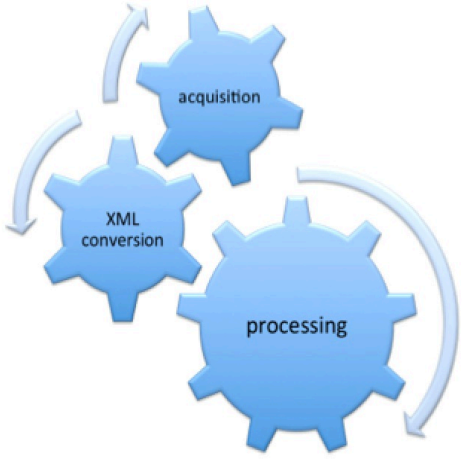
Search



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

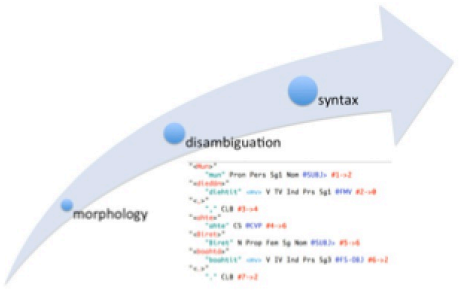

Autshumato ITE :: st_åm
Editor – st_åm.odt
Sametingets årsmelding omfatter Sametingets politiske og forvaltningsmessige aktivitet i 2011.
Årsmeldingen rapporterer om oppfølgingen av hovedmålene og delmålene for de enkelte fagområdene og bruken av virkemidlene i budsjettet for 2011.
Sametingets regnskap for 2011 er vedlagt årsmeldingen.
Sametingsvalget
Sámediggeválggat
Forskrift om valg til Sametinget ble endret i 2011.
<segment 0013>
Endringen innebærer at Sametingets valgmanntall også utarbeides i de årene det er kommune- og fylkestingsvalg.
Dette valgmanntallet 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 valgmanntall.
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 oastimiid láhka ja láhkaásahus rievdaduvvui 2001:s.
<44/44/60% /Users/ttr000/Documents/tm/nob2sme-2012-10-21/nob2sme-2012-10-21-000679.tmx >
2) Høring – Sametingsvalget – forskrift om valg til Sametinget
Sámediggeválga – láhkaásahus 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 –rievdaduvui jagi 2006.
<37/37/44% /Users/ttr000/Documents/tm/nob2sme-2012-10-
Glossary
forskrift = láhkaásahus
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

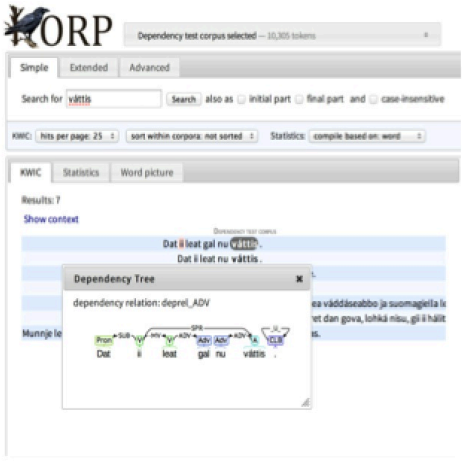


Lexicon

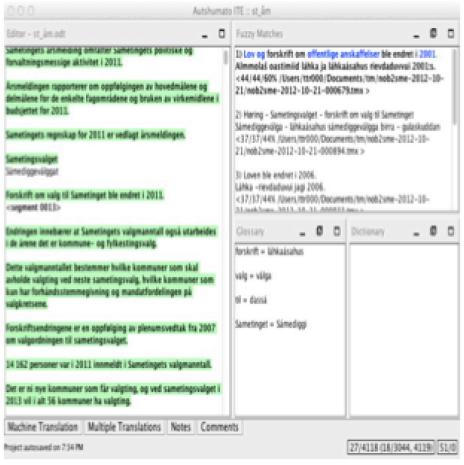
Analysis



Search



CAT



The aqcuired knowledge can be used for other Saami languages.

Bures boahtin!  Buerie båeteme!

Know-how

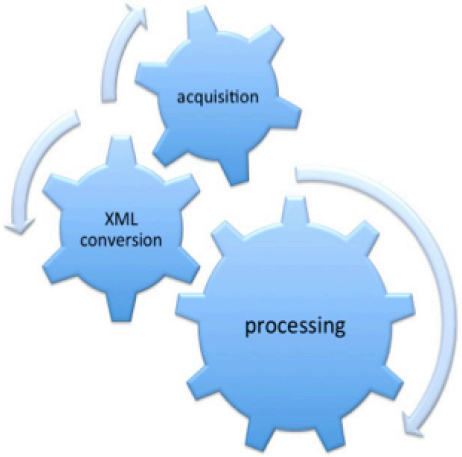
- Corpus
- Tools
- Pipelines

How?

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

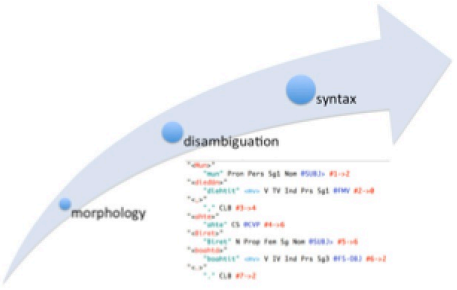
A good teamwork is essential to the final result.

Corpus

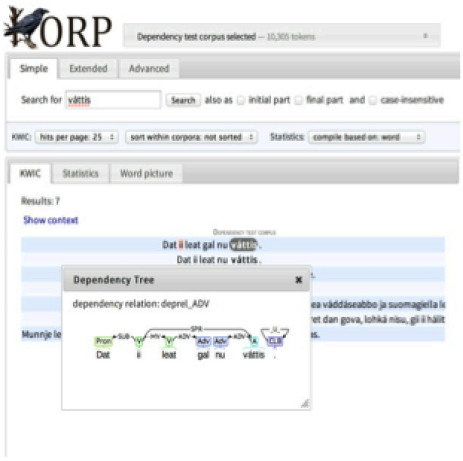


Lexicon

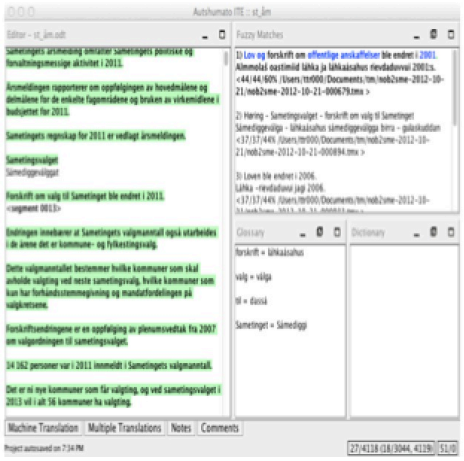
Analysis



Search



CAT



Share know-how