Apertium - open-source rule-based machine translation
Google Summer of Code

Jacob Nordfalk
DTU Diplom Campus Ballerup
jacno@dtu.dk / jacob.nordfalk@gmail.com

KUA 13. april 2014

Jacob Nordfalk

- Jacob Nordfalk
  - Assoc. professor (lektor) at Danish Technical University, Campus Ballerup
  - Java, Linux, Web, Android, advanced topics
  - Author of 3 Java programming books, http://javabog.dk
  - Active in the community of the International Language Esperanto
  - Self-chosen unemployed in Nepal 2006-2008
  - Working with i.a. Apertium and English-Esperanto
  - Mentoring i.a. Swedish-Danish, Norwegian-Danish and English-Esperanto during Google Summer of Code
- Android
  - Mentor of ~100 Android projects
  - A few apps on Google Play, i.a. DR Radio, DR Nyheder and Apertium offline translator

Approaches to Machine Translation

Rule-based

Deep/Shallow transfer

Some or full syntactical parsing

Apertium
Open source
Shallow transfer

Statistically based

No parsing

Uses parallel corpora to piece together a text in source language

Unreliable, but vivid translations
Lots of grammatical errors

Google Translate
Closed source

Google Translate (2010)

En tilfældig side (fra svensk wikipedia):

Trakterna kring Fredriksberg räknas som bebodda sedan 1600-talet.

Google Oversæt giver (nok via engelsk):

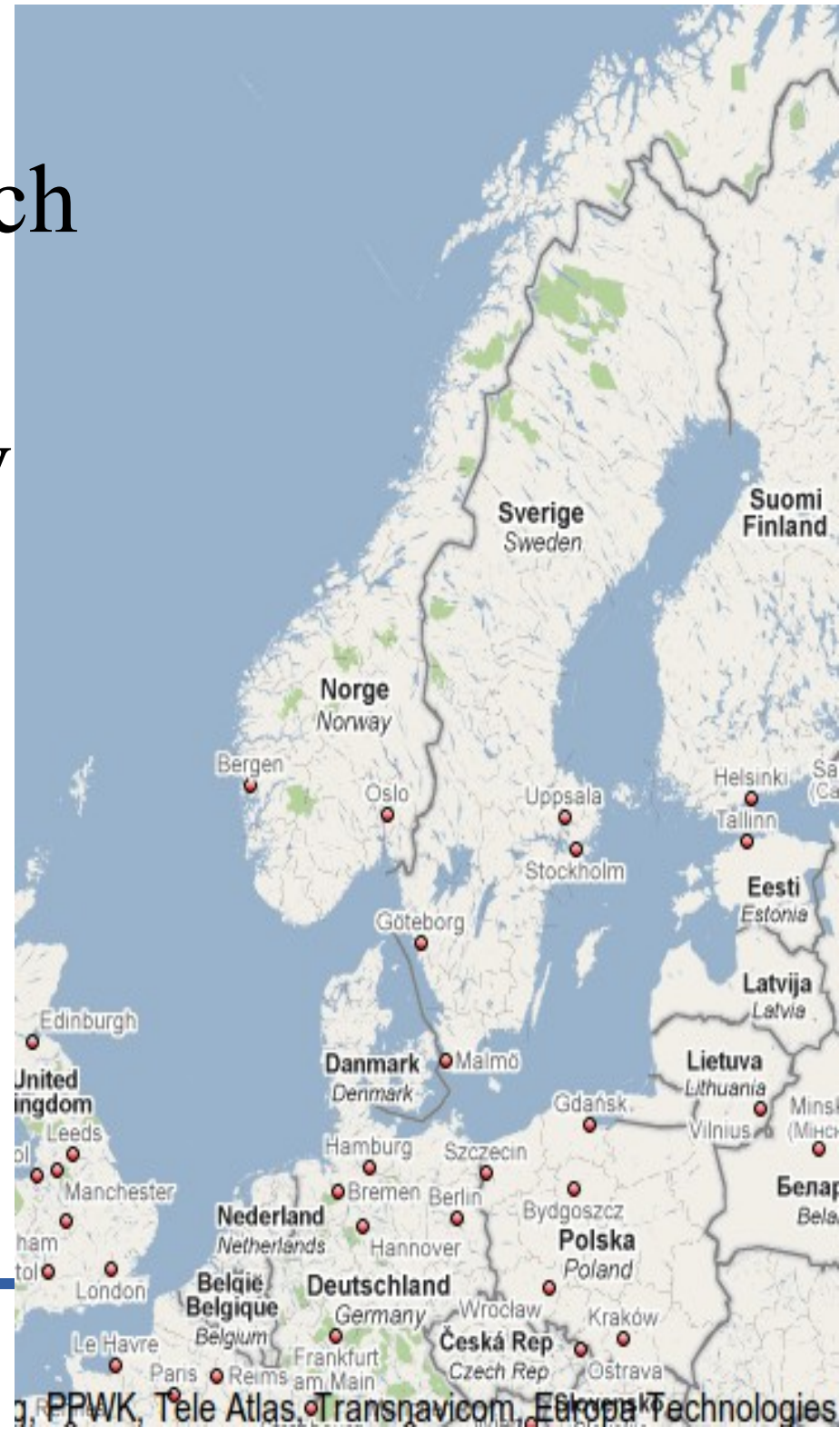
Områderne omkring Fredriksberg tælles som har været besat siden 1600-tallet.

(er Google nu egentlig en hjælp?)

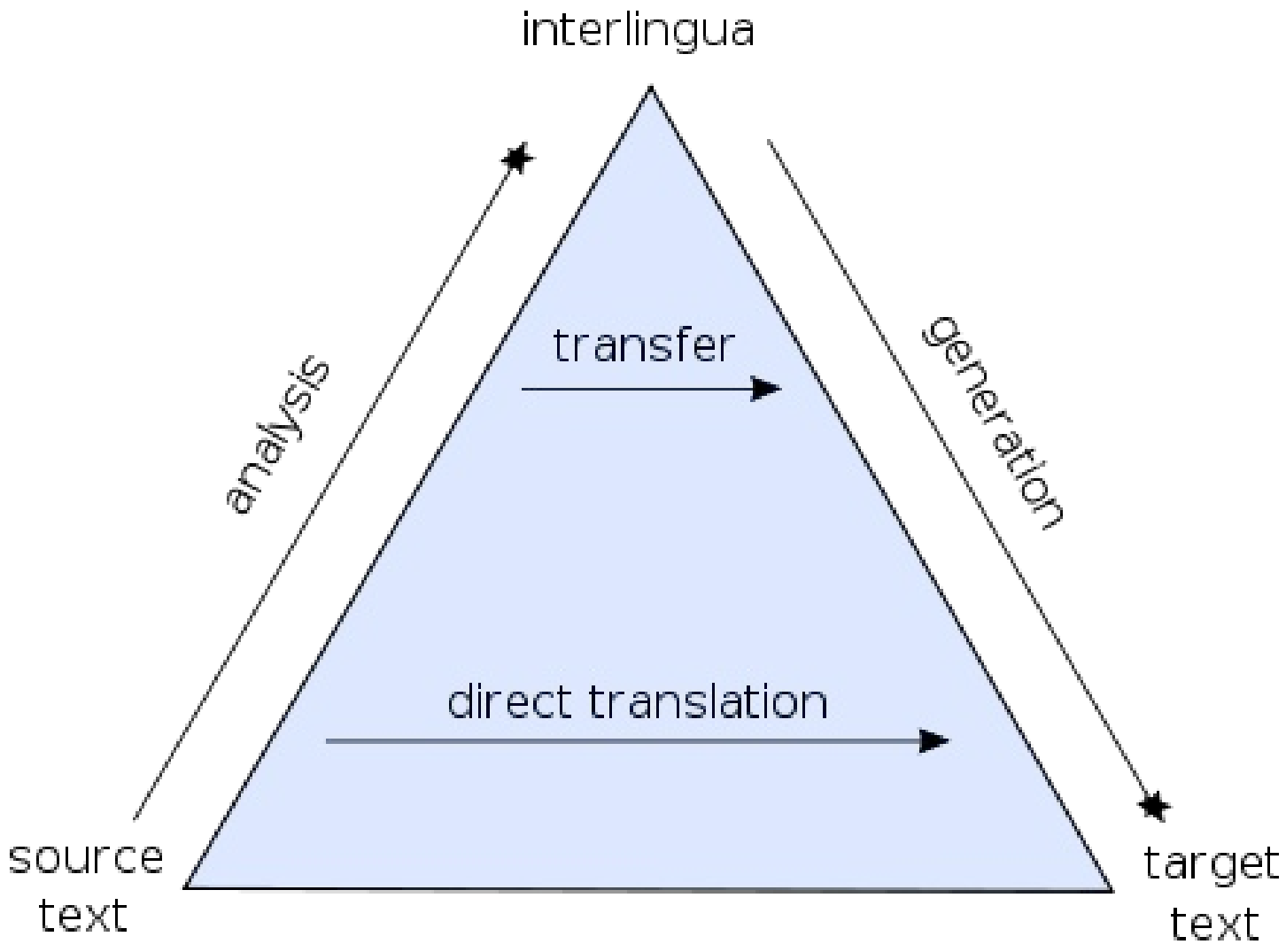
Se http://translationparty.com

Swedish and Danish

- Standardised in the 12th to 15th centuries out of the Old Norse which was spoken across Scandinavia.
- The languages are largely mutually intelligible
  - focus on production of text for dissemination (for post-editing)
  - production of text for assimilation (understanding) less important



Approaches to MT



Bernard Vauquois' pyramid

The Apertium project

Apertium is an open-source (GPL) machine translation platform. The platform provides

- a language-independent MT engine
- tools to manage linguistic data for language pairs
- linguistic data for >70 languages
  - Esperanto ⇆ English Swedish ⇆ Danish Catalan ⇆ Romanian Welsh ⇆ English English ⇆ Afrikaans English ⇆ Catalan English ⇆ Spanish English ⇆ Polish Esperanto ← Catalan Esperanto ← Spanish Esperanto←Nepali Spanish ⇆ Catalan Spanish ⇆ Galician Spanish ⇆ Italian Spanish ⇆ Portuguese Spanish ← Romanian Basque⇆Spanish French ⇆ Catalan French ⇆ Spanish Occitan ⇆ Catalan Occitan ⇆ Spanish Serbo-Croatian⇆Macedonian Nynorsk⇆Bokmål ...

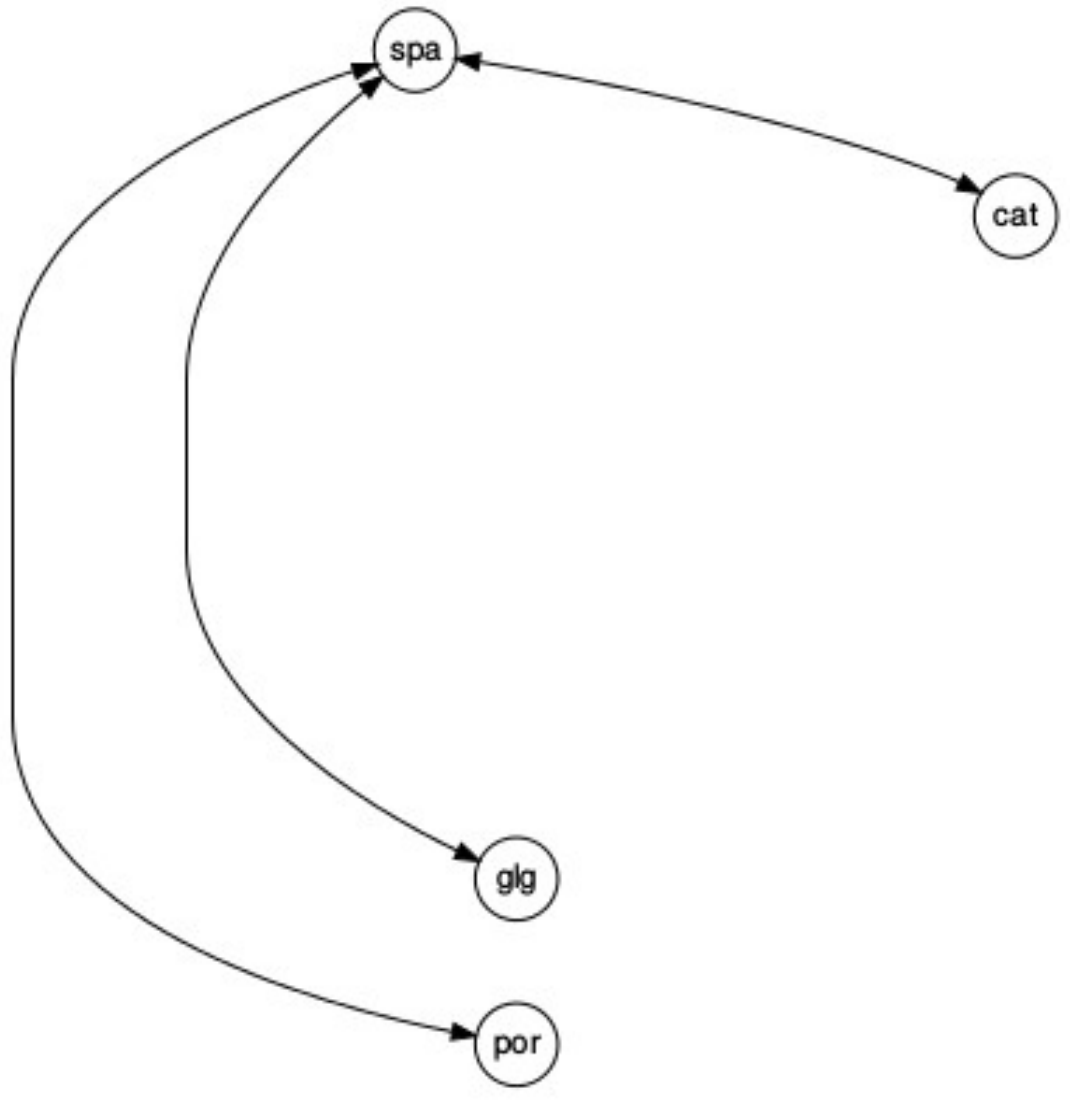
Apertium

- rule based
- open source (GPL - Gnu Public License)
  - the engine
  - all linguistic data
  - all supplementary tools
- host for an active and very helpfull community
  - Mailing list
  - Chat #apertium on Freenode
- relatively easy to learn, significant contributions can be made also by students without university degrees in linguistics or IT
- not that resource demanding
  (a language pair takes ~4 months to make)

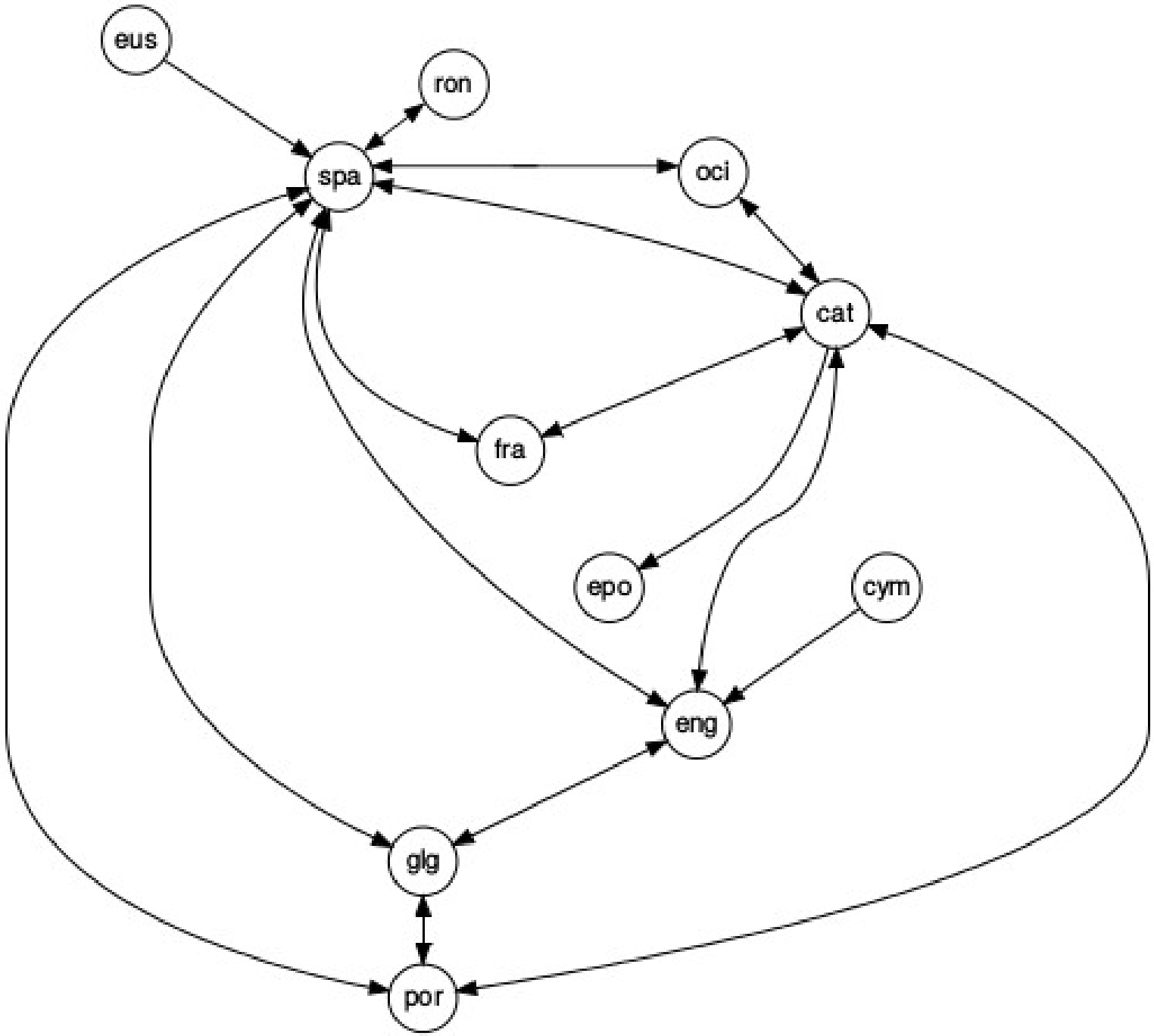
Apertium project

- Rule based "shallow-transfer" MT
  - finite-state transducers for all lexical processing operations
  - hidden Markov models for part-of-speech tagging
    - and/or Constraint Grammar
  - multi-stage finite-state based chunking for structural transfer
    - and/or a lot of other stuff :-)
- Processing happens in stages
  - Each stage i a seperate program
  - Output from one stage is input to the next stage

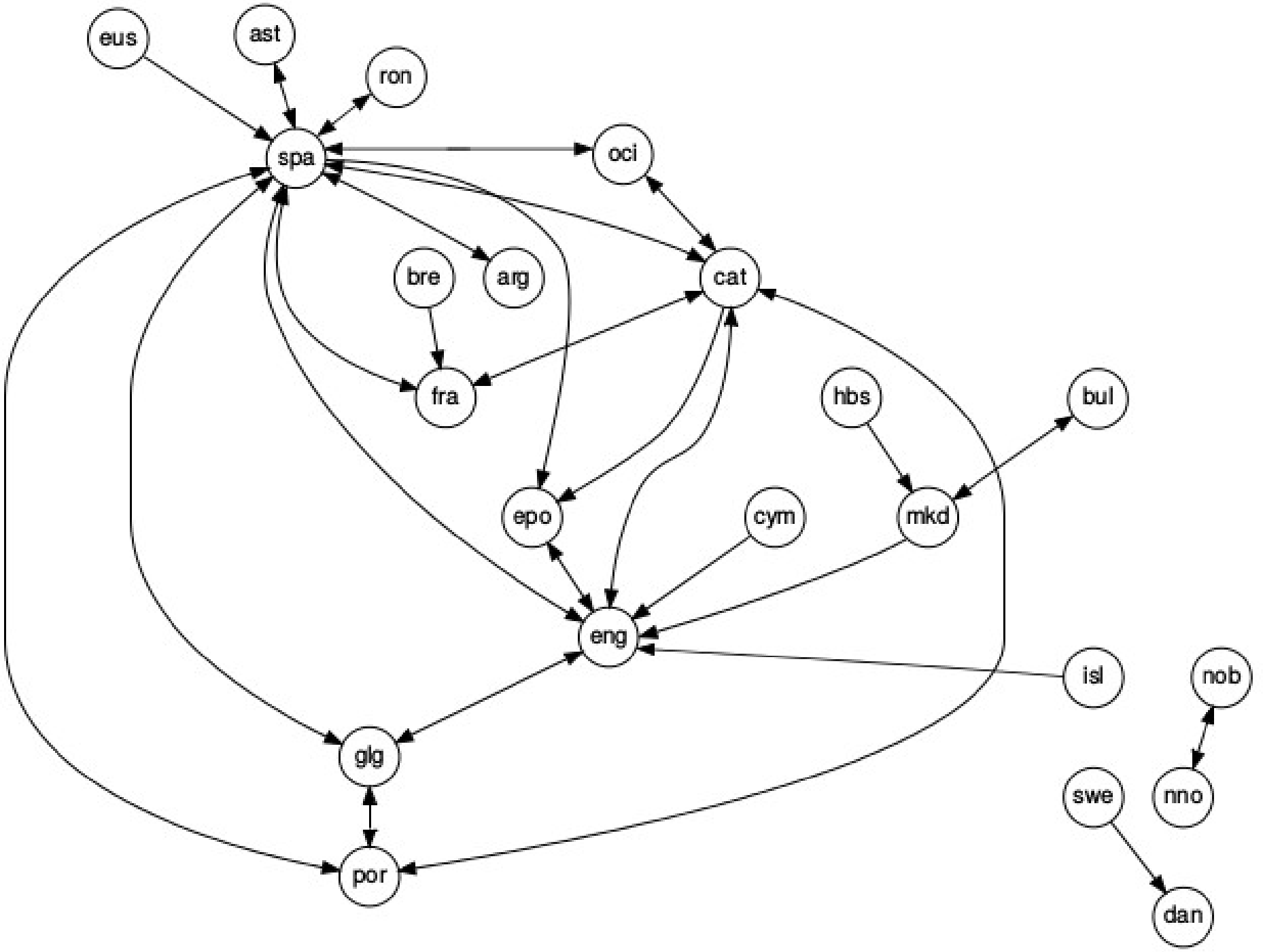
2005



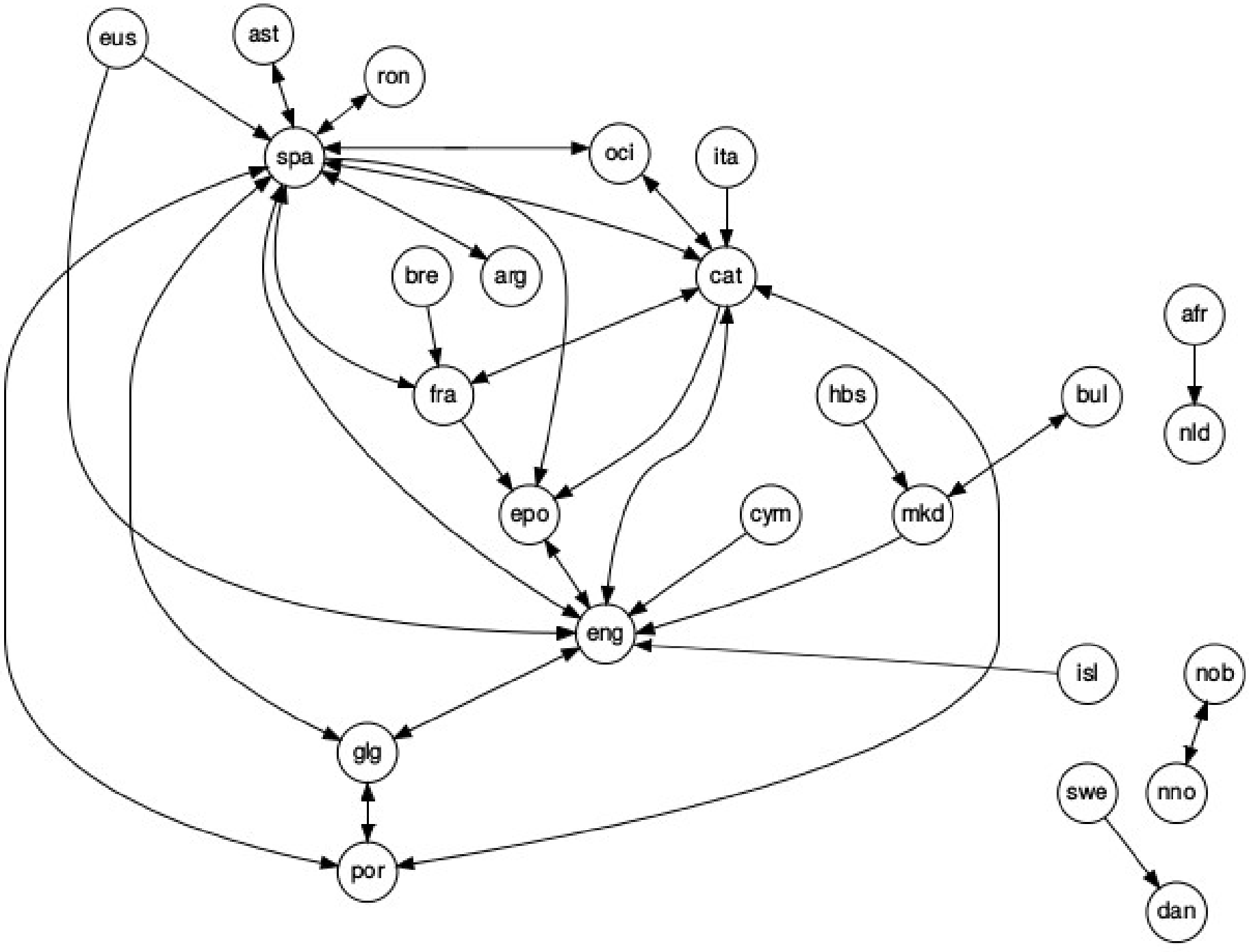
2008



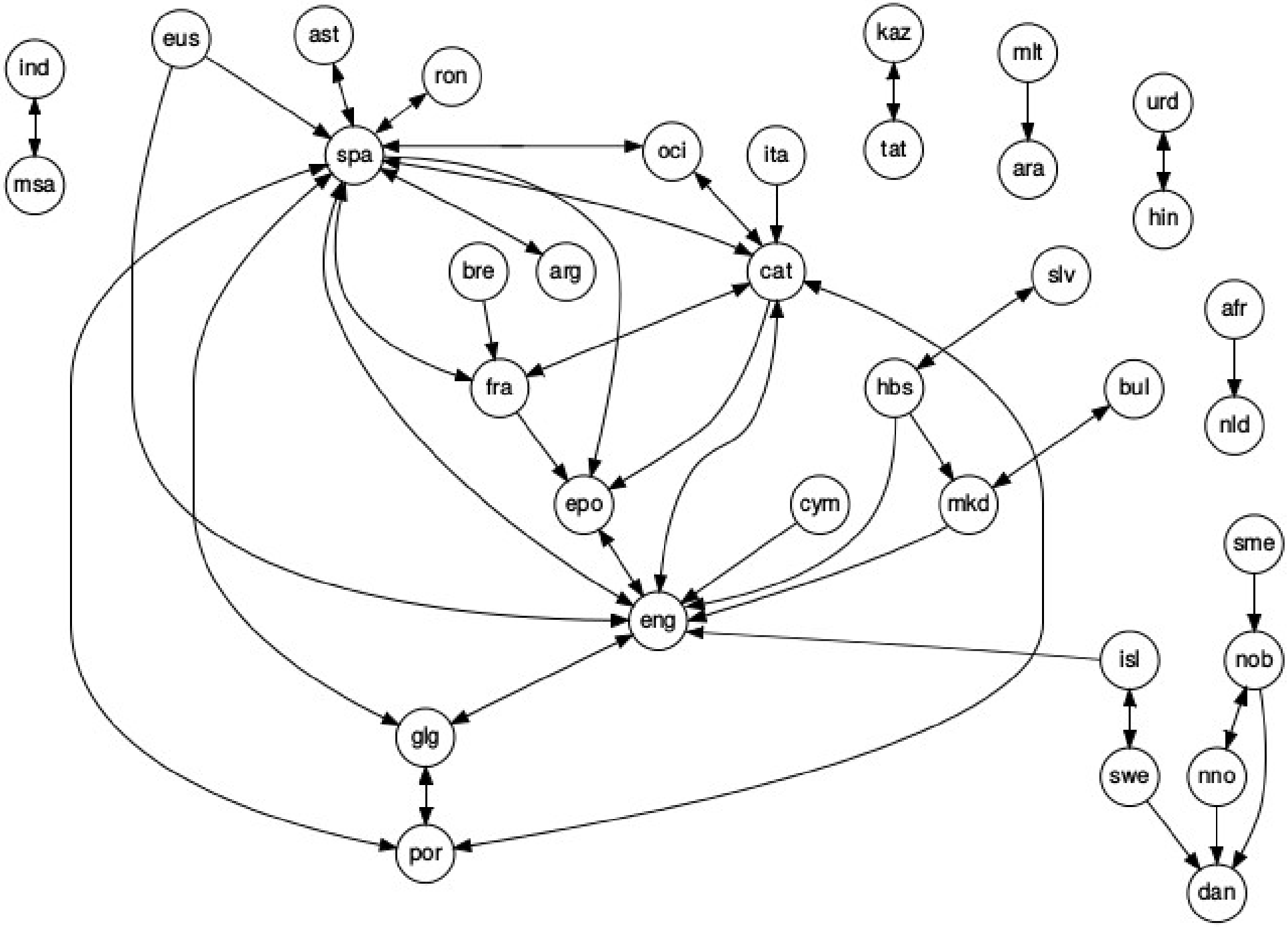
2010


eus
ast
ron
spa
oci
bre
arg
cat
fra
hbs
bul
epo
cym
mkd
eng
isl
nob
glg
swe
nno
por
dan

2012


eus
ast
ron
spa
oci
ita
bre
arg
cat
afr
fra
hbs
bul
nld
epo
cym
mkd
eng
isl
nob
glg
swe
nno
por
dan

2014

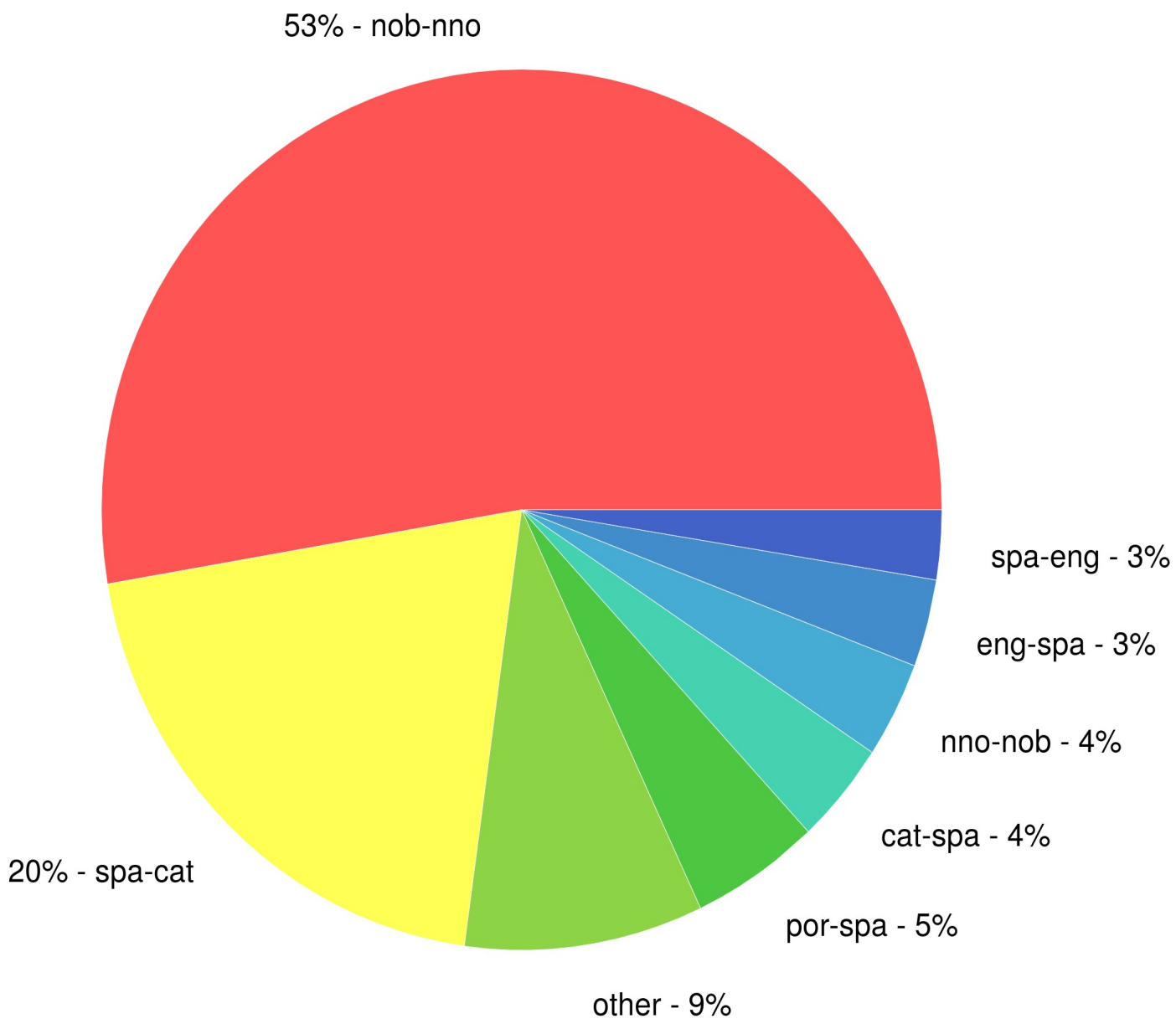

ind
msa
eus
ast
ron
spa
oci
ita
kaz
tat
mlt
ara
urd
hin
bre
arg
cat
slv
afr
fra
hbs
bul
nld
epo
cym
mkd
sme
eng
isl
nob
glg
swe
nno
por
dan

Status 2013 (not updated to '15)

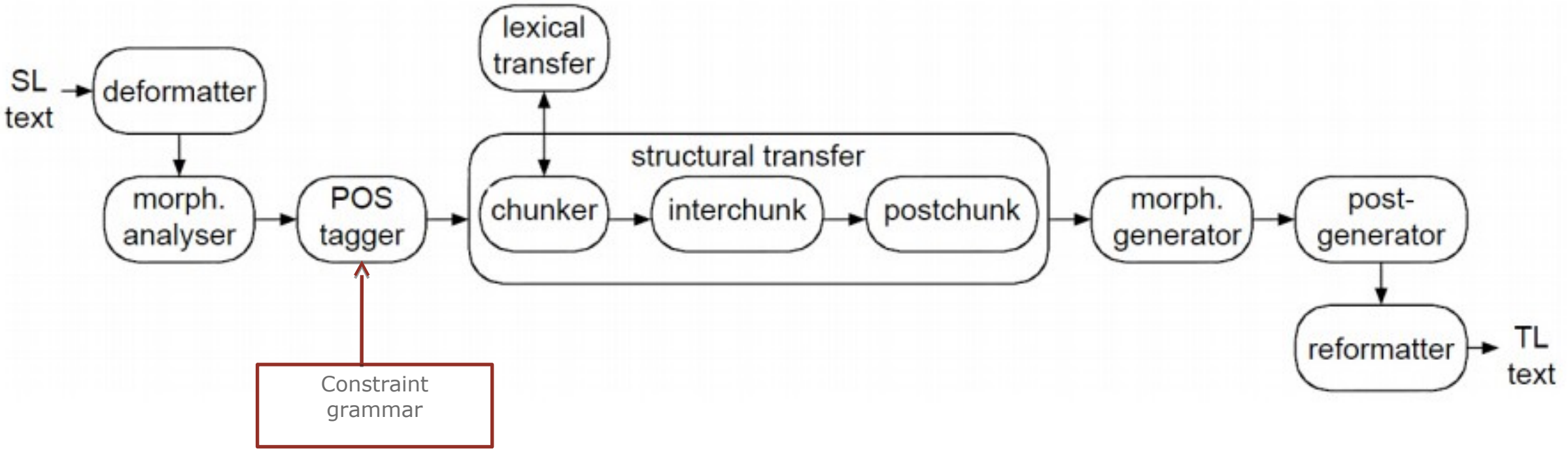
- Released
  - 35 language pairs released - ready to use
  - 8 language pairs 'staging' - 'need a little more work'
  - 30 language pairs in nursery - 'needs more work'
  - data for 129 pairs waiting in incubator

See http://wiki.apertium.org/wiki/List_of_language_pairs

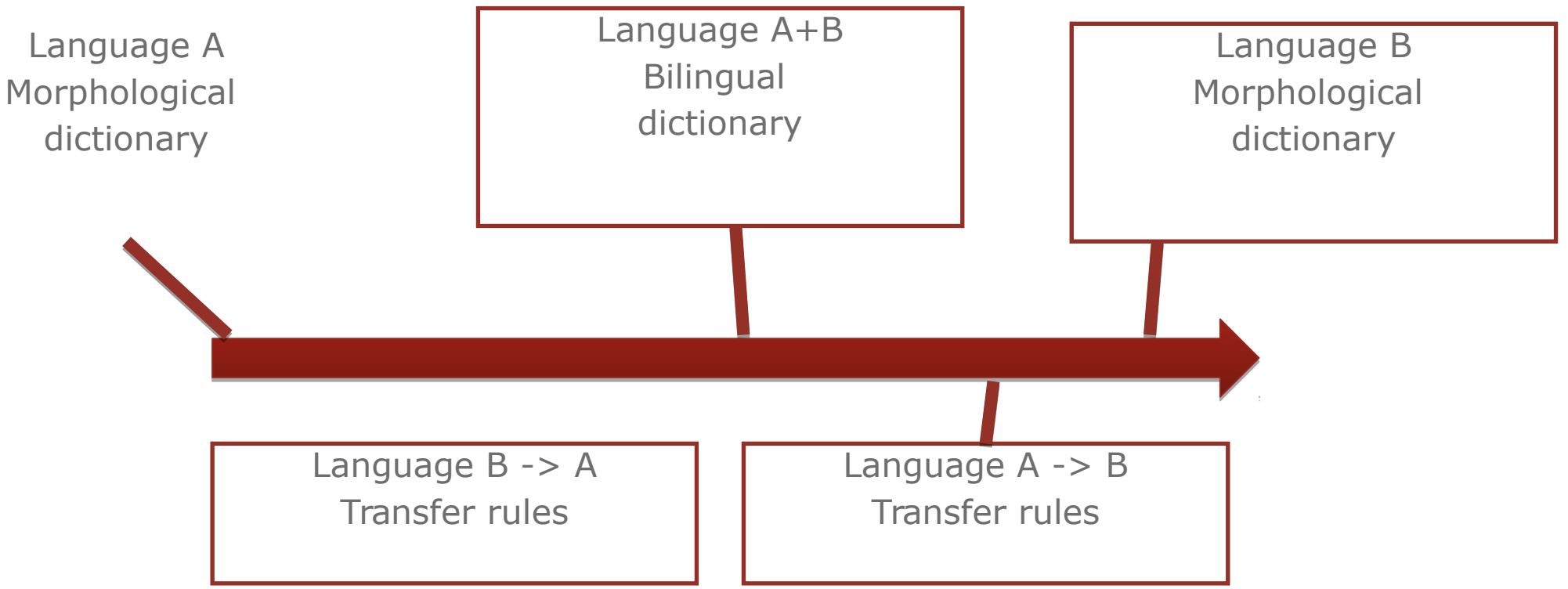
Usage stats may 2015



Architecture of Apertium MT



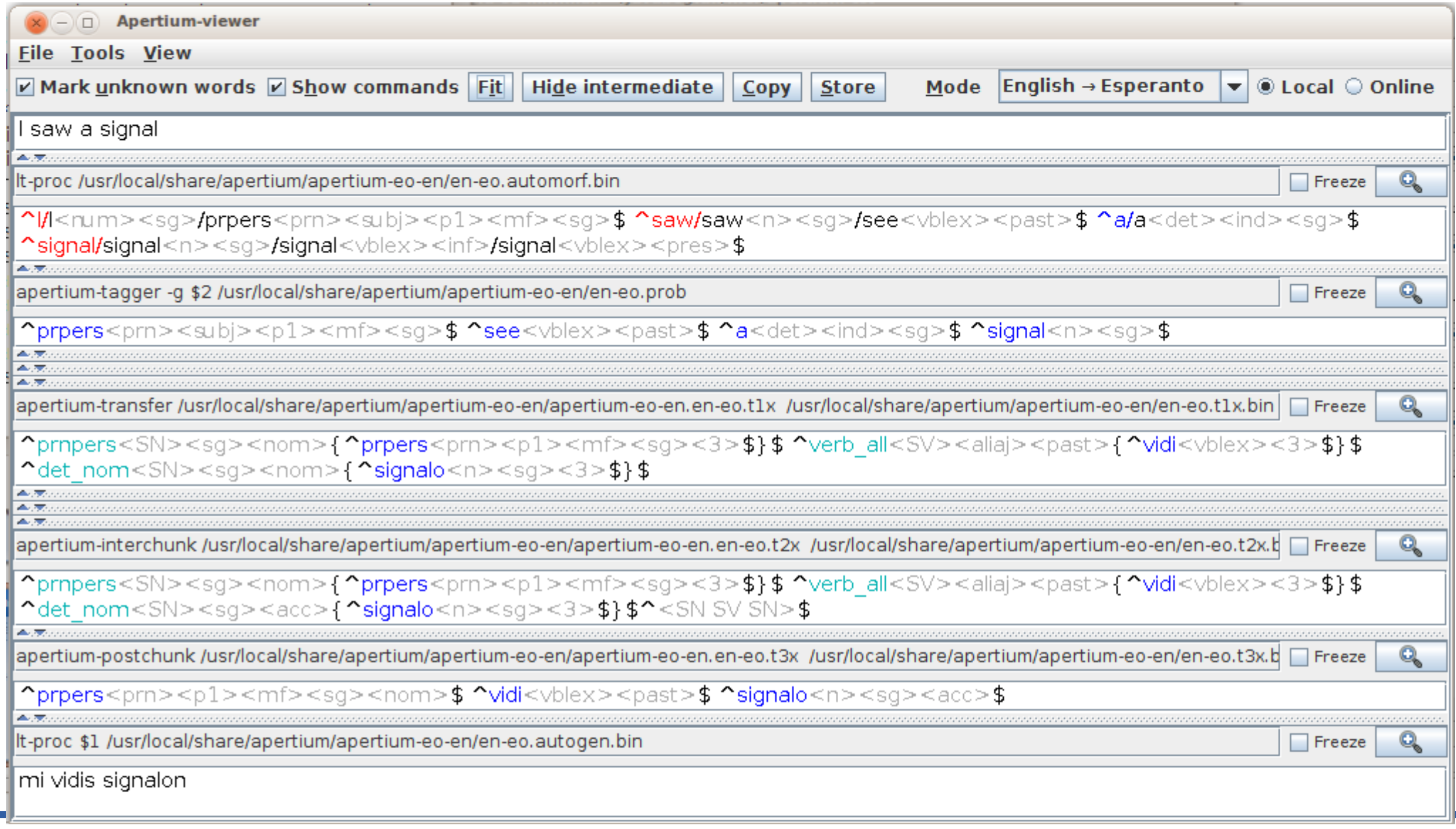
Dictionaries and transfer rules



Morfological analysis

- English and Esperanto language pair
  - Marking accusative (n)
    - I saw a signal -> Mi vidis signalo**n**
  - English words can be very ambigious
    - ^saw/
      - saw<n><sg>/
      - saw<vblex><inf>/
      - saw<vblex><pres>/
      - saw<vblex><imp>/
      - see<vblex><past>$

Apertium-viewer showing the pipeline



Dictionaries and rules

- The three dictionaries
  - English monodix

```
<e lm="see"><i>s</i><par n="s/ee__vblex"/></e>
<e lm="saw"><i>saw</i><par n="house__n"/></e>
<e lm="saw"><i>saw</i><par n="accept__vblex"/></e>
```

  - English-Esperanto bidix

```
<e><p><l>vidi<s n="vblex"/></l><r>see<s n="vblex"/></r></p></e>
<e><p><l>segilo<s n="n"/></l><r>saw<s n="n"/></r></p></e>
<e><p><l>segi<s n="vblex"/></l><r>saw<s n="vblex"/></r></p></e>
```

  - Esperanto monodix

```
<e lm="vidi"><i>vid</i><par n="verb__vblex"/></e>
<e lm="segilo"><i>segilo</i><par n="nom__n"/></e>
<e lm="segi"><i>seg</i><par n="verb__vblex"/></e>
```

- Transfer rules
  - Also written in XML

Paradigm definitions

```
<pardef n="s/ee__vblex">
  <e>       <p><l>ee</l>        <r>ee<s n="vblex"/><s n="inf"/></r></p></e>
  <e>       <p><l>ee</l>        <r>ee<s n="vblex"/><s n="pres"/></r></p></e>
  <e>       <p><l>ees</l>       <r>ee<s n="vblex"/><s n="pres"/><s n="p3"/><s n="sg"/></r></p></e>
  <e>       <p><l>aw</l>        <r>ee<s n="vblex"/><s n="past"/></r></p></e>
  <e>       <p><l>een</l>       <r>ee<s n="vblex"/><s n="pp"/></r></p></e>
  <e>       <p><l>eeing</l>     <r>ee<s n="vblex"/><s n="ger"/></r></p></e>
</pardef>

<pardef n="house__n">
  <e>       <p><l></l>          <r><s n="n"/><s n="sg"/></r></p></e>
  <e>       <p><l>s</l>         <r><s n="n"/><s n="pl"/></r></p></e>
</pardef>

<pardef n="accept__vblex">
  <e>       <p><l></l>          <r><s n="vblex"/><s n="inf"/></r></p></e>
  <e>       <p><l></l>          <r><s n="vblex"/><s n="pres"/></r></p></e>
  <e>       <p><l>s</l>         <r><s n="vblex"/><s n="pres"/><s n="p3"/><s n="sg"/></r></p></e>
  <e>       <p><l>ed</l>        <r><s n="vblex"/><s n="past"/></r></p></e>
  <e>       <p><l>ed</l>        <r><s n="vblex"/><s n="pp"/></r></p></e>
  <e>       <p><l>ing</l>       <r><s n="vblex"/><s n="ger"/></r></p></e>
</pardef>


<e lm="see"><i>s</i><par n="s/ee__vblex"/></e>
<e lm="saw"><i>saw</i><par n="house__n"/></e>
<e lm="saw"><i>saw</i><par n="accept__vblex"/></e>
...
<e lm="adsorbate">      <i>adsorbate</i><par n="house__n"/></e>
<e lm="adsorbent">      <i>adsorbent</i><par n="house__n"/></e>
<e lm="adsorption">     <i>adsorption</i><par n="house__n"/></e>
<e lm="adulation">      <i>adulation</i><par n="house__n"/></e>
<e lm="adult">          <i>adult</i><par n="house__n"/></e>
....
<e lm="appeal">         <i>appeal</i><par n="house__n"/></e>
<e lm="appeal">         <i>appeal</i><par n="accept__vblex"/></e>
<e lm="appear">         <i>appear</i><par n="accept__vblex"/></e>
<e lm="appearance">     <i>appearance</i><par n="house__n"/></e>
<e lm="appease">        <i>appeas</i><par n="liv/e__vblex"/></e>
<e lm="append">         <i>append</i><par n="accept__vblex"/></e>
```

```
see:see<vblex><inf>
see:see<vblex><pres>
sees:see<vblex><pres><p3><sg>
saw:see<vblex><past>
seen:see<vblex><pp>
seeing:see<vblex><ger>

saw:saw<n><sg>
saws:saw<n><pl>

saw:saw<vblex><imp>
saw:saw<vblex><inf>
saw:saw<vblex><pres>
saws:saw<vblex><pres><p3><sg>
sawed:saw<vblex><past>
sawing:saw<vblex><ger>
```

Exercises ;-)

- Download the presentations
  - http://javabog.dk/filer/apertium/
- Visit the wiki
  - http://wiki.apertium.org
- Try it out on your PC/Mac/Linux
  - http://wiki.apertium.org/wiki/Apertium-viewer
- Install the app
  - Search for 'Apertium' in Google Play
- Visit the IRC chat (http://wiki.apertium.org/wiki/IRC)
- På dansk
  http://wiki.apertium.org/wiki/Dansk_introduktion

Exercises for interested parties

- Install the source code
  - http://wiki.apertium.org/wiki/Installation
  - Easiest on (Ubuntu) Linux
  - Java source is only for execution
  - C++ source is for development and execution
- Get in touch (IRC is best for getting help)
  - http://wiki.apertium.org/wiki/Contact
-

Less related languages (chunking) / 3-stage transfer

I saw a signal

becomes after disambiguation

```
^prpers<prn><subj><p1><mf><sg>$
^see<vblex><past>$
^a<det><ind><sg>$
^signal<n><sg>$.
```

which is transfered and chunked into

```
^prnpers<SN><p1><mf><sg>{^prpers<prn><subj><2><3><4>$}$
^verb<SV><past>{^vidi<vblex><past>$}$
^nom<SN><sg><nom>{^signalo<n><2><3><4>$}$.
```

and transformed by rule SN SV SN -> SV SV SN<acc>

```
^prnpers<SN><p1><mf><sg>{^prpers<prn><subj><2><3><4>$}$
^verb<SV><past>{^vidi<vblex><past>$}$
^nom<SN><sg><acc>{^signalo<n><2><3><4>$}$.
```

and then unchunked

```
^prpers<prn><subj><p1><mf><sg>$
^vidi<vblex><past>$
^signalo<n><sg><acc>$.
```

Less related languages (chunking) / 3-stage transfer

- all kinds of operations can be applied
  - word order can be changed (verb in end of sentence)
  - gender, tense and other tags can be replaced or moved
  - words and special features of source/destination language can be removed/added

Swedish-Danish

Structural transfer sv-da

- Double definiteness

Den stora utmaningen (‘The big challege’)
^Den<det>**<def>**<ut><sg>$ ^stor<adj><pst><un><pl><ind>$ ^utmaning<n><ut><sg>**<def>**<nom>$
^Den<det>**<def>**<ut><sg>$ ^stor<adj><pst><un><pl><ind>$ ^udfordring<n><ut><sg>**<ind>**<nom>$
Den store udfordring

- Swedish supine verb form

Han hade blivit trott (‘He had been believed’)
^Han<prn><subj><p3><m><sg>$ ^**ha**<vbhaver><past><actv>$ ^bli<vblex>**<supn>**<actv>$
^tro<vblex><pp><nt><sg><ind>$
^Han<prn><subj><p3><m><sg>$ ^**være**<vbser><past><actv>$ ^blive<vblex>**<pp>**$ ^tro<vblex><pp>$
Han var blevet troet
(sometimes the auxillary verb is omitted in Swedish - *Han blivit trott*. This is currently not supported)

- Changes in auxiliary verbs

Två personer har börjat (‘Two people has begun’)
^Två<num><un><pl>$ ^person<n><ut><pl><ind><nom>$ ^ha<vbhaver><pres><actv>$
^börja<vblex><supn><actv>$
^To<num><un><pl>$ ^person<n><ut><pl><ind><nom>$ ^være<vbser><pres><actv>$ ^begynde<vblex><pp>$
To personer er begyndt (‘Two people is begun’)

Structural transfer sv-da

- Changes in present passive formation

Det publiceras ('It is being published')
^Det<prn><subj><p3><nt><sg>$ ^publicera<vblex><pres><pasv>$
^Det<prn><subj><p3><nt><sg>$ ^publicere<vblex><pres><pasv>$
Det publiceres

Det upprepas ('It is being repeated')
^Det<prn><subj><p3><nt><sg>$ ^upprepa<vblex>**<pres><pasv>**$
^Det<prn><subj><p3><nt><sg>$ ^**blive<vblex><pres><actv>**$ ^gentage<vblex>**<pp>**$
Det bliver gentaget

Changes in past passive formation

Det publicerades ('It was being published')
^Det<prn><subj><p3><nt><sg>$ ^publicera<vblex><past><pasv>$
^Det<prn><subj><p3><nt><sg>$ ^blive<vblex><past><actv>$ ^publicere<vblex><pp>$
Det blev publiceret

Det upprepades ('It was being repeated')
^Det<prn><subj><p3><nt><sg>$ ^upprepa<vblex><past><pasv>$
^Det<prn><subj><p3><nt><sg>$ ^blive<vblex><past><actv>$ ^gentage<vblex><pp>$
Det blev gentaget

Challenges in transfer

- Gender and number change in determiners, adjective, nouns
  - <nt> (Neuter), <ut> (Common) ⇆
    <un> (Common/Neuter), <GD> (gender to be determined)
  - <sg>, <pl> ⇆
    <sp>, <ND> (number to be determined)
  - Concordance: gender, number of determiner and adjectives follow must noun
  - Synthetic adjectives (better, best vs. more good, most good)

Bidix paradigms for simplicity

- <sp> words (singular and plural have same form)
  ^datum/datum<n><nt**><sp>**<ind><nom>$ →
  ^dato/dato<n><ut>**<sg>**<ind><nom>$ or
  ^datoer/dato<n><ut>**<pl>**<ind><nom>$

En atlas		^atlas<n><ut><sg><ind><nom>$		^atlas<n><nt><sp><ind><nom>$		Et atlas
Atlasen		^Atlas<n><ut><sg><def><nom>$		^Atlas<n><nt><sg><def><nom>$		Atlasset
Två atlaser	→	^atlas<n><ut><pl><ind><nom>$	→	^atlas<n><nt>**<sp>**<ind><nom>$	→	To atlas
De två atlasen		^atlas<n><ut><pl><def><nom>$		^atlas<n><nt>**<sp>**<ind><nom>$		De to atlas

```
<pardef n="sgpl_sp__n">
  <e r="RL"><p><l><s n="ND"/><s n="ind"/></l><r><s n="sp"/><s n="ind"/></r></p></e>
  <e r="LR"><p><l><s n="sg"/><s n="ind"/></l><r><s n="sp"/><s n="ind"/></r></p></e>
  <e r="LR"><p><l><s n="pl"/><s n="ind"/></l><r><s n="sp"/><s n="ind"/></r></p></e>
  <e>       <p><l><s n="sg"/><s n="def"/></l><r><s n="sg"/><s n="def"/></r></p></e>
  <e>       <p><l><s n="pl"/><s n="def"/></l><r><s n="pl"/><s n="def"/></r></p></e>
</pardef>

<e><p><l>atlas<s n="n"/><s n="ut"/></l><r>atlas<s n="n"/><s n="nt"/></r></p><par n="sgpl_sp__n"/></e>

<e><p><l>datum<s n="n"/><s n="nt"/></l><r>dato<s n="n"/><s n="ut"/></r></p><par n="sp_sgpl__n"/></e>
```

Dictionary entries for adjectives

- Swedish monodix

```
<pardef n="aktiv__adj">
<e><p><l></l>    <r><s n="adj"/><s n="pst"/><s n="ut"/><s n="sg"/><s n="ind"/></r></p></e>
<e><p><l>t</l>   <r><s n="adj"/><s n="pst"/><s n="nt"/><s n="sg"/><s n="ind"/></r></p></e>
<e><p><l>e</l>   <r><s n="adj"/><s n="pst"/><s n="m"/><s n="sg"/><s n="def"/></r></p></e>
<e><p><l>a</l>   <r><s n="adj"/><s n="pst"/><s n="un"/><s n="pl"/><s n="ind"/></r></p></e>
<e><p><l>a</l>   <r><s n="adj"/><s n="pst"/><s n="un"/><s n="sp"/><s n="def"/></r></p></e>

<e><p><l>are</l> <r><s n="adj"/><s n="comp"/><s n="un"/><s n="sp"/></r></p></e>
<e><p><l>ast</l> <r><s n="adj"/><s n="sup"/><s n="un"/><s n="sp"/><s n="ind"/></r></p></e>
<e><p><l>aste</l><r><s n="adj"/><s n="sup"/><s n="un"/><s n="sp"/><s n="def"/></r></p></e>
</pardef>

<e lm="vit">              <i>vit</i><par n="aktiv__adj"/></e>
```

- Swedish-Danish bidix

```
<e><p><l>vit<s n="adj"/></l><r>hvid<s n="adj"/></r></p><par n="aktiv_aktiv__adj"/></e>
```

- Danish monodix

```
<pardef n="aktiv__adj">
<e><p><l></l>   <r><s n="adj"/><s n="pst"/><s n="ut"/><s n="sg"/><s n="ind"/></r></p></e>
<e><p><l>t</l>  <r><s n="adj"/><s n="pst"/><s n="nt"/><s n="sg"/><s n="ind"/></r></p></e>
<e><p><l>e</l>  <r><s n="adj"/><s n="pst"/><s n="un"/><s n="pl"/><s n="ind"/></r></p></e>
<e><p><l>e</l>  <r><s n="adj"/><s n="pst"/><s n="un"/><s n="sp"/><s n="def"/></r></p></e>
</pardef>

<e lm="hvid">             <i>hvid</i><par n="aktiv__adj"/></e>
```

Bidix paradigms... for simplicity (?)

- Adjective follows gender, number and can be synthetic

En vit atlas.		^vit<adj><pst><ut><sg><ind>$	^hvid<adj><pst><nt><sg><ind>$	Et hvidt atlas.
Atlasen			^mere<preadv>$	Atlasset
Två vitare atlaser	→	^vit<adj><comp><un><sp>$	→ ^hvid<adj><pst><un><pl><ind>$	→ To mere hvide atlas
De två vitaste atlaserna		^vit<adj><sup><un><sp><def>$	^mest<preadv>$	De to mest #hvid
			^hvid<adj><pst><sup><nt><pl><ind><def>$	atlassene

```
<pardef n="aktiv_aktiv__adj">
<e>       <p><l><s n="pst"/><s n="un"/><s n="sp"/><s n="def"/></l><r><s n="pst"/><s n="un"/><s n="sp"/><s n="def"/></r></p></e>
<e>       <p><l><s n="pst"/><s n="un"/><s n="pl"/><s n="ind"/></l><r><s n="pst"/><s n="un"/><s n="pl"/><s n="ind"/></r></p></e>

<e r="LR"><p><l><s n="pst"/><s n="m"/><s n="sg"/><s n="def"/></l><r><s n="pst"/><s n="un"/><s n="sp"/><s n="def"/></r></p></e>
<e r="LR"><p><l><s n="pst"/><s n="ut"/></l><r><s n="pst"/><s n="ut"/></r></p></e>
<e r="LR"><p><l><s n="pst"/><s n="nt"/></l><r><s n="pst"/><s n="nt"/></r></p></e>

<e r="RL"><p><l><s n="pst"/><s n="GD"/></l><r><s n="pst"/><s n="ut"/></r></p></e>
<e r="RL"><p><l><s n="pst"/><s n="GD"/></l><r><s n="pst"/><s n="nt"/></r></p></e>

<e r="LR"><p><l><s n="comp"/><s n="un"/><s n="sp"/></l><r><s n="unsint"/><s n="comp"/><s n="GD"/><s n="ND"/></r></p></e>
<e r="RL"><p><l><s n="sint"/><s n="comp"/><s n="un"/></l><r><s n="comp"/><s n="ut"/></r></p></e>
<e r="RL"><p><l><s n="sint"/><s n="comp"/><s n="un"/></l><r><s n="comp"/><s n="nt"/></r></p></e>

<e r="LR"><p><l><s n="sup"/><s n="un"/><s n="sp"/></l><r><s n="unsint"/><s n="sup"/><s n="GD"/><s n="ND"/></r></p></e>
<e r="RL"><p><l><s n="sint"/><s n="sup"/><s n="un"/></l><r><s n="sup"/><s n="ut"/></r></p></e>
<e r="RL"><p><l><s n="sint"/><s n="sup"/><s n="un"/></l><r><s n="sup"/><s n="nt"/></r></p></e>
</pardef>

<e>       <p><l>vit<s n="adj"/></l>                    <r>hvid<s n="adj"/></r></p><par n="aktiv_aktiv__adj"/></e>
```

Evaluation

	Number entries
Monolingual dict. (sv)	5,230 lemmas
Bilingual dict.	6,854 lemmas
Monolingual dict. (da)	10,694 lemmas
Transfer rules (sv → da)	17 rules

System	Edit distance	WER	PWER
Apertium	350	30	28
Gramtrans	304	26	20
Google	415	35	22

Sv original: Historik.
Da postedit: Historik.
Apertium : Historik.
Gramtrans : Historik.
Gooogle SMT: Historie.

Sv: Trakterna kring Fredriksberg räknas som bebodda sedan 1600-talet.
Da: Områderne omkring Fredriksberg regnes som beboede siden 1600-tallet.
Ap: *Trakterna omkring *Fredriksberg regnes som *bebodda siden 1600-talen.
Gr: Områderne omkring Fredriksberg regnes som beboede siden 1600-talet.
Go: Områderne omkring Fredriksberg tælles som har været besat siden 1600-tallet.

Sv: Området kring Fredriksberg utgjorde ursprungligen den södra delen av Nås finnmark,
Da: Området omkring Fredriksberg udgjorde oprindelig den sydlige del af Nås finnmark,
Ap: Området omkring *Fredriksberg *utgjorde oprindeligt den *södra delen af Nås *finnmark,
Gr: Området omkring Fredriksberg udgjorde oprindeligt den sydlige del af Nås finnmark,
Go: Området omkring Frederiksberg var oprindeligt den sydlige del af Reachable Sverige,

Corpus	Running tokens	Known tokens	Coverage
Wikipedia	30,662,861	22,030,690	71.84%
EuroParl	15,531,107	12,499,971	80.48%

Sv: och området räknas som en del av Västerdalarna
Da: og området regnes som en del af Västerdalarna
Ap: og området regnes som en del af *Västerdalarna
Gr: og området regnes som en del af Västerdalarna
Go: og området regnes som en del af den vestlige del af Dalarna

Sv: (till skillnad från övriga Ludvika kommun, som räknas till Bergslagen).
Da: (til forskel fra øvrige Ludvika kommune, som regnes til Bergslagen).
Ap: (til forskel fra øvrige *Ludvika kommune, som regnes til *Bergslagen).
Gr: (til forskel fra den øvrige Ludvika kommune, som regnes til Bergslagen).
Go: (i modsætning til andre Ludvika Kommune, som rækker_Bergslagen).

Evaluation

System	Edit distance	WER	PWER
Apertium	350	30	28
Gramtrans	304	26	20
Google	415	35	22

	Translation	Gloss
Original	*Det finns en kort överfart vid det baltiska havet vid Helsingborg, på vilket ställe Själland kan ses från Skåne, ett vanligt tillhåll för vikingar.*	There exists a short passage by the Baltic Sea by Helsingborg, on which place Sjælland can be seen from Skåne, a common hangout for Vikings.
Apertium	Det findes en kort *överfart* ved det *baltiska* havet ved Helsingborg, på hvilket *ställe Själland* kan ses fra Skåne, et *vanligt* tilhold før vikinger.	It exists a short *överfart* by the *baltiska* Sea by Helsinborg, on which *ställe Själland* can be seen from Skåne, a *vanligt* order before Vikings.
Gramtrans	Der findes en kort overfart ved det baltiske hav ved Helsingborg, på hvilket sted *Själland* kan ses fra Skåne, et sædvanligt tilhold for vikinger.	There exists a short passage by the Baltic Sea by Helsingborg, on which place *Själland* can be seen from Skåne, a common order for Vikings.
Google	Der er en kort *passage* i Østersøen i Helsingborg, i hvilken plads *Zealand* kan ses fra *Scania*, en regelmæssig tilholdssted for vikingerne.	There is a short *passage* in the Baltic Sea in Helsingborg, in which space/place/seat *Zealand* can be seen from *Scania*, a regular hangout for the Vikings.

Table 4: Comparison of the three systems for a single sentence. Unknown words are marked with *emphasis* and incorrect translations are underlined.

Licence soup

- This presentation may be distributed under the terms of the GNU GPL, GNU FDL and CC-BY-SA licences.
  - GNU GPL v. 3.0
    - http://www.gnu.org/licenses/gpl.html
  - GNU FDL v. 1.2
    - http://www.gnu.org/licenses/gfdl.html
  - CC-BY-SA v. 3.0
    - http://creativecommons.org/licenses/by-sa/3.0/

Norvegian-Danish

Translation Challenge

1 Hvor er James?

2 James og Mary er i hagen. Det er fint vær i dag, det er veldig varmt. I går var det veldig kaldt. De kunne ikke leke ute da. James og Mary elsker å leke, de leker alltid sammen i hagen utenfor det store huset.

3 James er en liten gutt og han er seks år gammel. Den lille piken er hans søster, hun er fem år gammel. James har en liten hund, den er også i hagen nå. Hunden liker å leke med de to barna. Hunden er veldig glad nå.

4 Har Mary også en hund? Nei, Mary har ikke noen hund, hun har en katt. Katten er i huset, den sover

5 .....

IRC

```
 begiak  apertium: spectre360 * 43810: /nursery/apertium-da-nb/dev/testvoc/: inconsistency-
         summary.sh, testvoc-summary.nob-dan.txt: update to work with bc
spectie  jonasfromseier, so run that
spectie  then try:
spectie
spectie  $ cat /tmp/nob-dan.testvoc | grep '<ij>'
spectie  ja ------>  ^ja<ij>$ ^.<sent><clb>$ ------>  ^ja<ij>$  ------>  ja
spectie  nei ------>  ^nei<ij>$ ^.<sent><clb>$ ------>  ^nej<ij>$  ------>  nej
spectie  jo ------>  ^jo<ij>$ ^.<sent><clb>$ ------>  ^@jo<ij>$  ------>  \@jo\<ij\>
spectie
spectie  and note:
spectie  ========================================
spectie  POS Total Clean With @ With # Clean %
spectie  ij 3 2 1 0 66.67
spectie  so if you add
spectie  <e><p><l>jo<s n="ij"/></l>          <r>jo<s n="ij"/></r></p></e>
spectie
spectie  to the bilingual dict
spectie  then you'll have cleaned interjection
spectie  s
         Disconnected
```

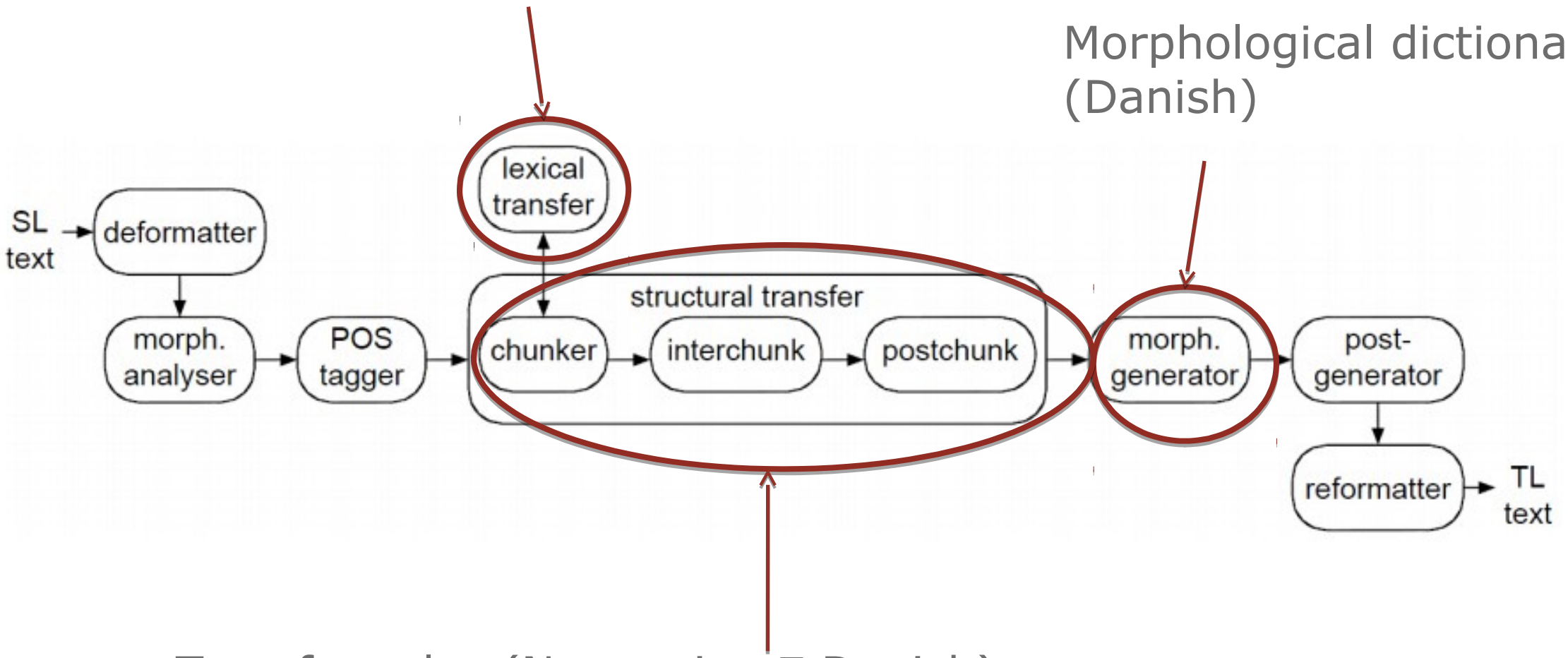
Work plan

week	plan
week 1	Identification of Norwegian-Danish morphology and syntax differences with the aid of grammar books (Norsk Referansgrammatik and Dansk for Norsker), supplementing with the study of parallel nb-da corpora and Norwegian texts.
week 2	Design and preliminary testing of transfer rules
week 3	Test and debugging of transfer rules.
week 4	Rectification of transfer rules erros.
Deliverable # 1	Complete set of transfer rules for the nb-da direction.
week 5	Commence work on Constraint Grammar. Porting of Oslo-Bergen tagger.
week 6	Design of Constraint Grammar.
week 7	Testing of Constraint Grammar.
Deliverable # 2	nb-da pair with Danish Constraint Grammar
week 8	Extension of bilingual dictionary and monodices to include the nynorsk variant of Norwegian.
week 9	Testvoc
week 10	Testvoc
week 11	Debugging transfer rules and CG.
week 12	Cleanup and dissemination

What did I work on?

Bilingual dictionary (Danish-Norwegian)

Morphological dictionary (Danish)



Transfer rules (Norwegian Danish)

Freely available resources

- Apertium Norwegian Nynorsk-Bokmål system
- Newspaper corpora (Språkbanken's Gold Standard Corpus)
- Frequency lists (The Oslo Corpus of Tagged Norwegian Texts)
- Online Nynorsk-Bokmål dictionary (Oslo University/Språkrådet)
- Oslo-Bergen Tagger

Major differences

- Double definiteness

  Den fine damen -> *Den fine dame*

  "The posh lady"

  DET.DEF posh lady.DEF DET.DEF posh lady.IND

- Possessive pronoun placement + noun definiteness

  Moren sin -> *sin mor*

  "her mother"

  mother-DEF POSS POSS mother

- No disambiguation between existential *there* and formal subject *it*

  Det er en mand -> *Der er en mand / det er en mand*

  "It is a man" / "There is a man"

Major differences

- Double definiteness

  Den fine damen -> *Den fine dame*

  "The posh lady"

  DET.DEF posh lady.DEF DET.DEF posh lady.IND

- Possessive pronoun placement + noun definiteness

  Moren sin -> *sin mor*

  "her mother"

  mother-DEF POSS POSS mother

- ~~No disambiguation between existential *there* and formal subject *it*~~

~~*Det er en mand* -> *Der er en mand / det er en mand*~~

~~"It is a man" / "There is a man"~~

```
<rule comment="RULE DET (DEF) + NOUN (DEF) >  DET (DEF)  NOUN (IND)  den manden > den mand" >
  <pattern>
    <pattern-item n="det-dem"/>
    <pattern-item n="nom"/>
  </pattern>
  <action>

    <choose>
      <when><test><and>
        <equal><clip pos="2" side="sl" part="defnes"/><lit-tag v="def"/></equal>
        <equal><clip pos="1" side="sl" part="a_det"/><lit-tag v="det.dem"/></equal></and></test>
      <let><clip pos="2" side="tl" part="defnes"/><lit-tag v="ind"/> </let> </when>
    </choose>

    <choose>
      <when><test><equal>

        <!--if the noun has no tag for case, give it the case 'nom'-->
        <clip pos="2" side="sl" part="cas"/><lit v=""/></equal></test>
    <let><var n="kasus"/><lit-tag v="nom"/></let></when>

    <otherwise><let><var n="kasus"/><clip pos="2" side="sl" part="cas"/></let></otherwise>
</choose>

<call-macro n="determiner">
  <with-param pos="1"/>
</call-macro>

<out>
  <lu>
    <clip pos="1" side="tl" part="lemh"/>
    <clip pos="1" side="tl" part="a_det"/>
    <var n="køn"/>
    <clip pos="1" side="tl" part="nbr"/>
    <clip pos="1" side="tl" part="lemq"/>
  </lu>
  <b pos="1"/>
  <lu>
    <clip pos="2" side="tl" part="lemh"/>
```

Double definiteness rule patterns

Den beskatningen
DET NOUN

Den virksomhedsbeskatningen
DET CMP-NOUN

Den storvirksomhedsbeskatningen
DET CMP-CMP-NOUN

Den hårde beskatningen
DET ADJ NOUN

Den hårde årlige beskatningen
DET ADJ ADJ NOUN

Den hårde årlige virksomhedsbeskatningen
DET ADJ ADJ CMP-NOUN
...

Other rules

Gender

- Norwegian has three genders
  m / f / nt

- Luckily corresponds to
  m/f -> utrum
  nt -> nt

99.99% of the time!

Dictionaries

Bidix

```
<e><p><l>mærkelig<s n="adj"/></l>     <r>rar<s n="adj"/></r></p></e>
```

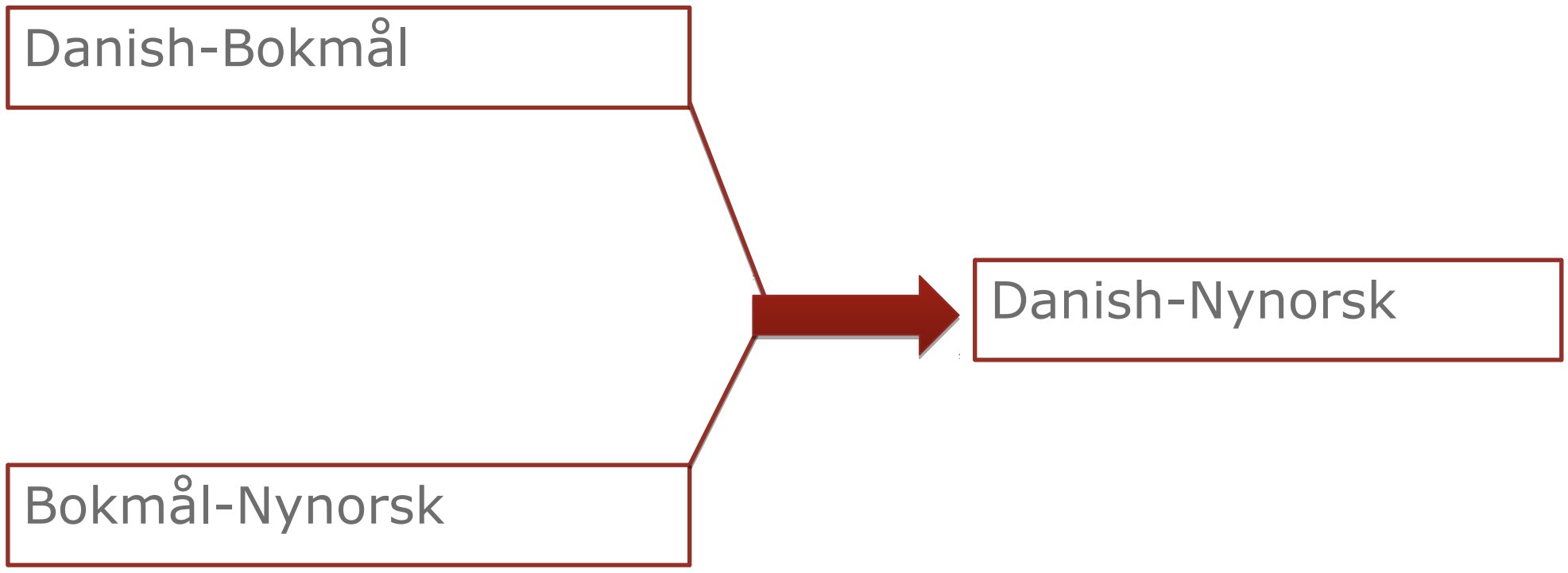
Danish monodix

```
<e lm="mærkelig">      <i>mærkelig</i><par n="lykkelig__adj"/></e>
```

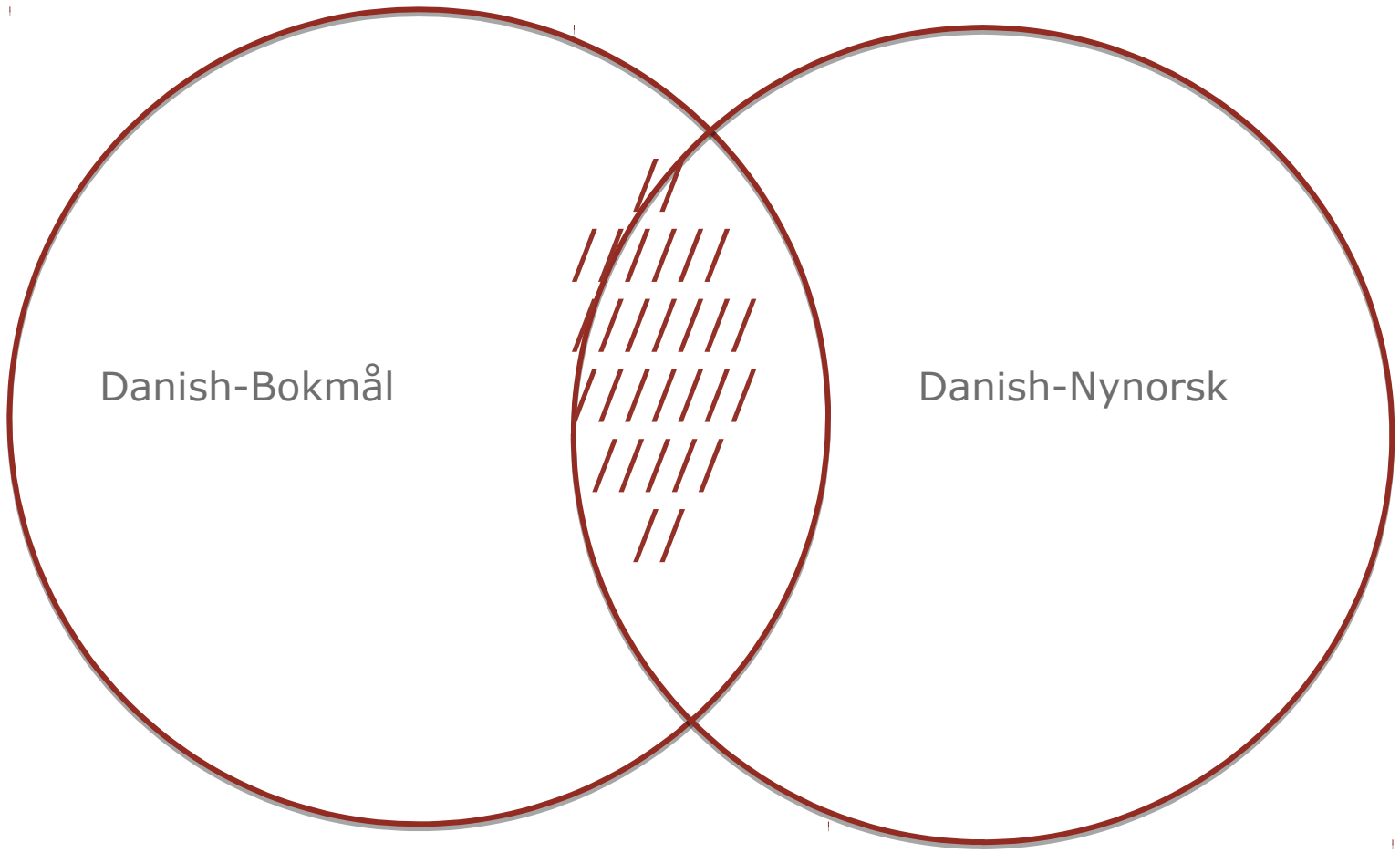
Paradigm definition in Danish monodix

```
<pardef n="lykkelig__adj”>

<e><p><l>   </l><r><s n="adj"/><s n="pst"/><s n="ut"/><s n="sg"/><s n="ind"/></r></p></e>
<e><p><l>t  </l><r><s n="adj"/><s n="pst"/><s n="nt"/><s n="sg"/><s n="ind"/></r></p></e>
<e><p><l>ere</l><r><s n="adj"/><s n="comp"/><s n="un"/><s n="sp"/></r></p></e>
<e><p><l>st </l><r><s n="adj"/><s n="sup"/><s n="un"/><s n="sp"/><s n="ind"/></r></p></e>
<e><p><l>ste</l><r><s n="adj"/><s n="sup"/><s n="un"/><s n="sp"/><s n="def"/></r></p></e>


</pardef>
```

Creation of Danish-Nynorsk



Intersection of Danish-Nynorsk and Danish-Bokmål



Tagging of bokmål/nynorsk in bidix

```
<e alt="nob"><p><l>mundtlig<s n="adj"/></l>
                <r>muntlig<s n="adj"/></r></p></e>

<e alt="nno"><p><l>mundtlig<s n="adj"/></l>
                 <r>munnleg<s n="adj"/></r></p></e>
```

Evaluation of Bokmål Danish

- 500 word text
- Run through MT system
- Translations individually post-edited by hand
- Script measures the edit distance etc. between the raw translation and the post-edited one

Overview [edit]

Tool	Word Error Rate	Edit distance	Position-independent correct words	Position-independent word error rate (PER)
Apertium	7,83 %	44	521	7,30 %
Gramtrans	7,12 %	40	528	6,05 %
Google Translate	28,67 %	160	460	22,04%

Evaluation of Bokmål Danish

Apertium

Mere end 4,3 millioner af disse findes i den engelsksproglige udgave.

Den næst største udgaven er den hollandske udgave med mere end 1,6 millioner artikler.

Der efter følger den *tyskspråklige udgaven med mere end 1,6 millioner artikler.

Konceptet Wikipedia er baseret på, er ikke nyt

Google Translate

Mere end 4,3 millioner af dem er på engelsk udgave

Den næststørste spørgsmål er den hollandske udgave med mere end 1,6 millioner artikler.

Derefter følger det tyske sprog Problem med mere end 1,6 millioner artikler.

Konceptet er baseret på Wikipedia , er ikke ny

Thanks

Google Summer of Code

Mentors:
Jacob Nordfalk
Francis Tyers
Kevin Unhammer

Line Burholt Kristensen for arranging this

Fagrådet for Lingvistik

Jessie for baking!