Shallow-transfer rule-based machine translation for Swedish to Danish

Francis M. Tyers

Dept. Lleng. i Sist. Informàtics,

Universitat d'Alacant,

Alacant. E-03070

ftyers@dlsi.ua.es

Jacob Nordfalk

Center for Videreuddannelse

Ingeniørhøjskolen i København

Denmark

jano@ihk.dk

Agenda

Apertium

Swedish-Danish

Language differences / structural transfer

Dictionary structure / lexical transfer challenges

Challenges in a Google Summer of Code (GSOC) project

Tools used to collect data

Evaluation

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 a lot of language pairs

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 ...

The Apertium project

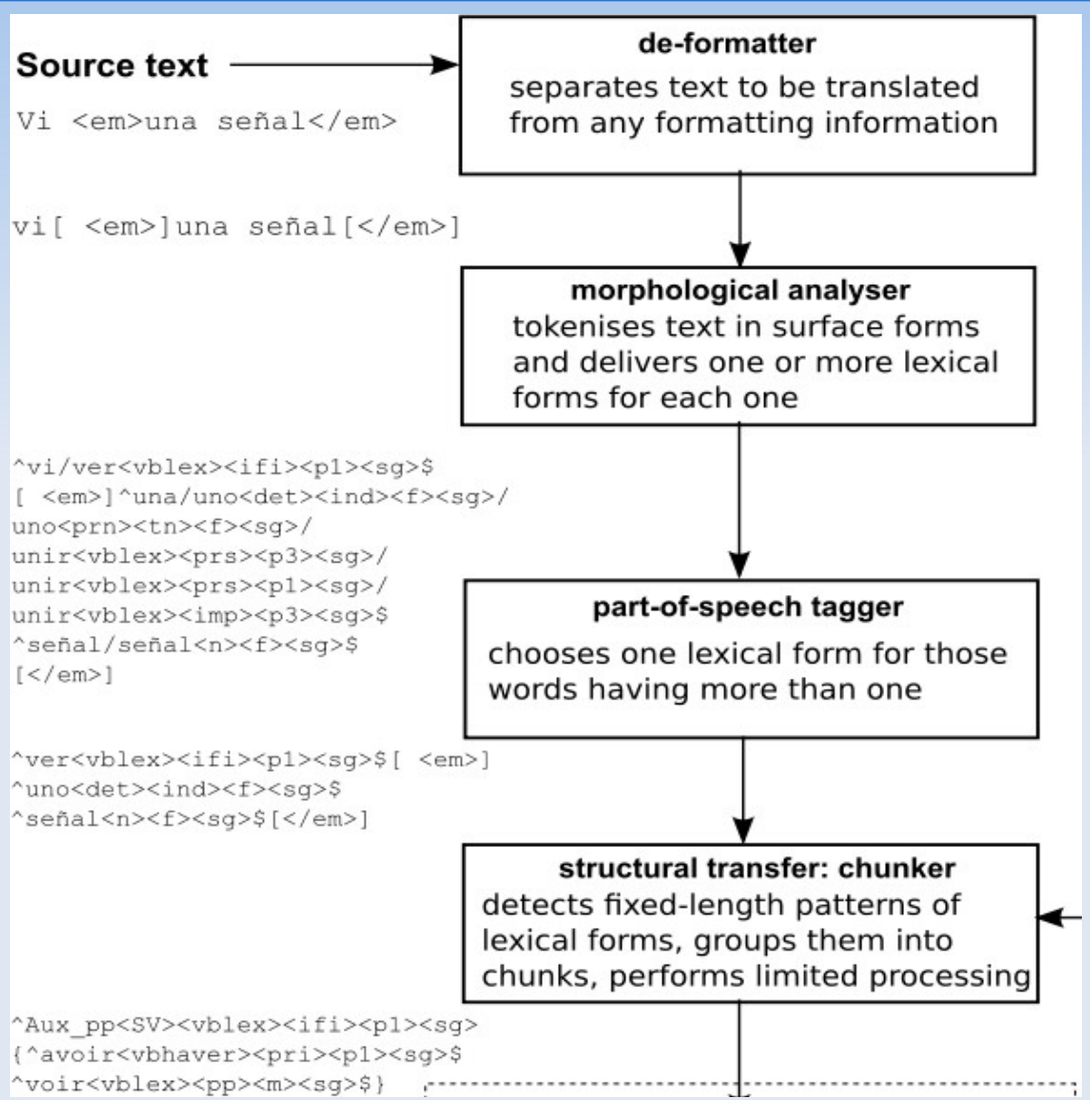
uses a shallow-transfer MT

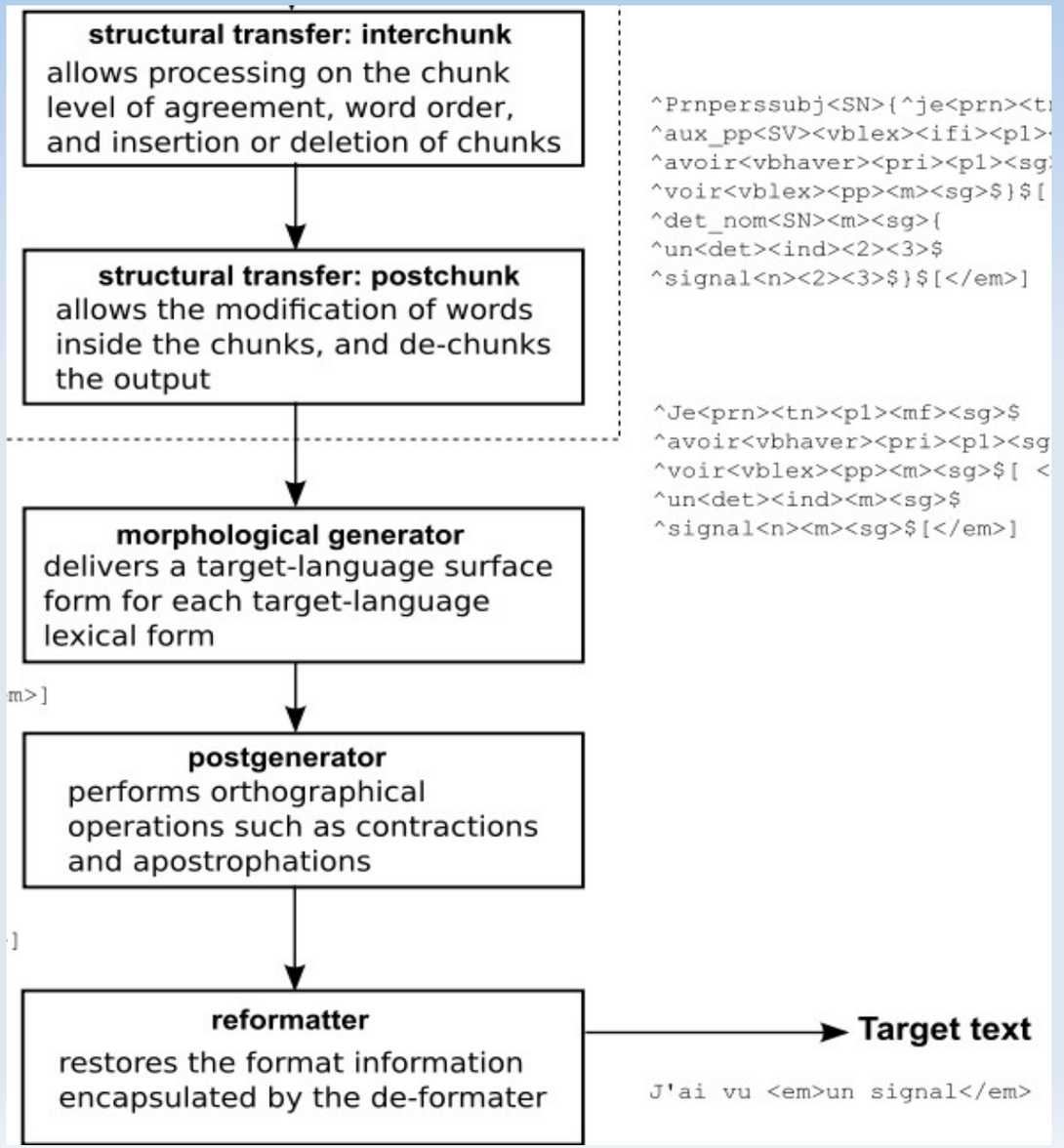
processes in stages, as in an assembly line:

de-formatting, morphological analysis, part-of-speech disambiguation (tagging), shallow structural transfer, lexical transfer, morphological generation, and re-formatting.

uses finite-state transducers for all lexical processing operations, hidden Markov models for part-of-speech tagging, and multi-stage finite-state based chunking for structural transfer.

Architecture of Apertium MT





Swedish and Danish

Standardised in the 12th to 15th centuries out of the Old Norse which was spoken across Scandinavia.

Swedish on the speech around Stockholm,

Danish on the speech around Copenhagen

The languages are largely mutually intelligible

- focus on production of text for dissemination (for post-editing)
- production of text for assimilation (understanding) less important

The people

(in order of amount of work with sv-da)

- Michael Kristensen
  - Google Summer of Code student of Apertium
- Francis M. Tyers
  - Dept. Lleng. i Sist. Informàtics, Universitat d'Alacant
- Jacob Nordfalk
  - Assoc. professor in Ingeniørhøjskolen i København / Copenhagen University College of Engineering, http://ihk.dk
  - Author of 3 Java programming books, http://javabog.dk
  - Active in the International Language Esperanto community, thanks to Fran & eo-es and eo-ca sponsored ABC Enciklopedioj, an active developer of Apertium Esperanto ⇆ English
  - GSoC mentor of Michael (officially, at least)

Structural transfer

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

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>
```

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		^me re<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>
```

Challenges of GSoC project

- First a lot of fun extracting data and doing cleanup scripts
  - Dictionaries started out big
  - shrank to ~5000 words and ~80% coverage
  - stil quite big to manage by manual cleaning
- A lot of design questions were left undecided
  - gender, number, case (nominative/genitive), active/passive
  - lots of testvoc problems which the student couldn't solve alone
  - Only in August 7th the Apertium and linguistic expert (Francis) met in person with the language expert (Michael) and the design was decided and main part of testvoc problems cleared
- GSoC student had lost all enthusiasm at that time :-(

Language resources used

Sorry,

international slide speed limit
of 60 slides/h would be exceeded !

(see the paper)

Evaluation

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

	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: 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.

What next

- Already now Apertium is usable for dissemination
- Upcoming work
  - Find new maintainer of sv-da
  - Cleanup da dix (comming tomorrow)
  - Increase dix coverage
  - The usual stuff (improve transfer, improve tagger etc)

Acknowledgements

Development was funded as part of the Google Summer of Code programme.

Many thanks to CSoC student Michael Kristensen for his big work.

Thanks to Thyge Larsen for assistance with post-edition and evaluation.

Licence soup