SMT within MOLTO's hybrid translation system

Cristina España-Bonet

Universitat Politècnica de Catalunya, TALP Research Center

–GF Summer School–

Barcelona, August 25th, 2011

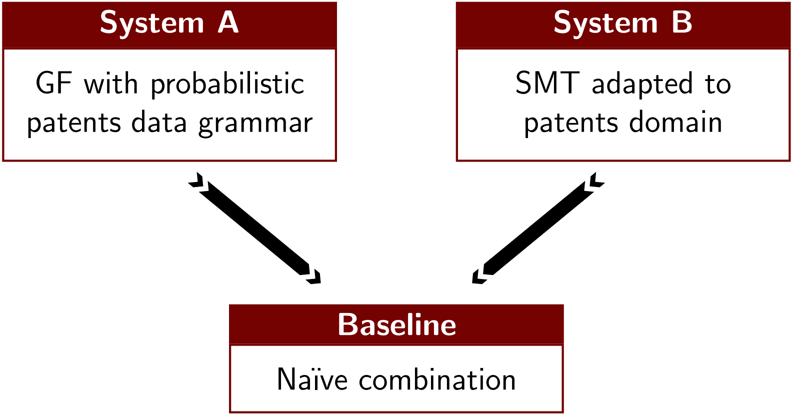
SMT within MOLTO's hybrid translation system

Overview

1. General view
2. Baselines
3. Hybrid systems
4. Conclusions

General view

Hybridisation: Baseline systems



System A

GF with probabilistic patents data grammar

System B

SMT adapted to patents domain

Baseline

Naïve combination

Baselines

Work on Baselines: GF –as explained by Ramona & Adam–

GF System

- **Parse**
- Apply patents **grammar**
- **Linearise**

Patents grammar

- **General** structure grammar
- **Compounds** grammar

Baselines

Work on Baselines: SMT

SMT baseline, Standard In-Domain System

- **Language model**: 5-gram interpolated Kneser-Ney discounting, SRILM Toolkit
- **Alignments**: GIZA++ Toolkit
- **Translation model**: Moses package
- **Weights optimization**: MERT against BLEU
- **Decoder**: Moses
- **Evaluation**: Asiya

Baselines

SMT baseline, Corpus

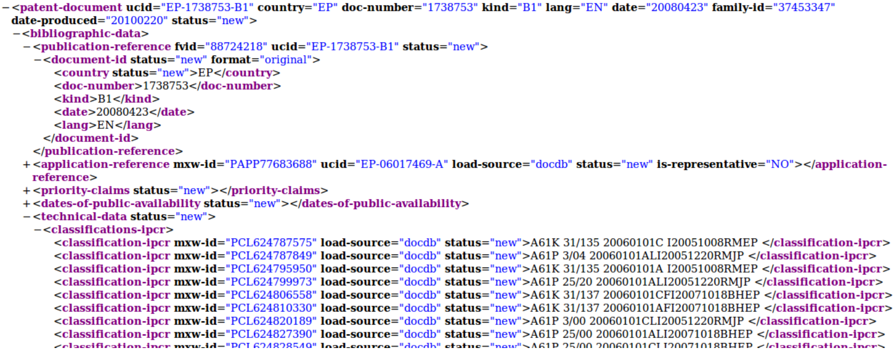
CLEF-IP 2010 Collection

Extract of the MAREC dataset, containing over 2.6 million patent documents pertaining to 1.3 milion patents from the EPO with some content in English, German and French.

Baselines

A Patent document

Patent document, **IPC** classification.

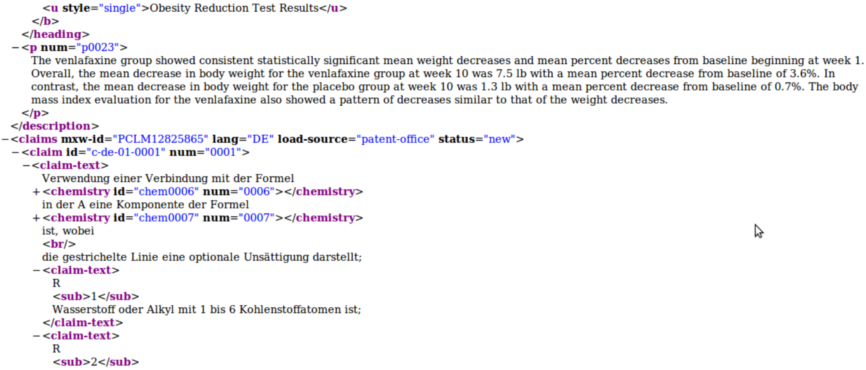


```
–<patent-document ucid="EP-1738753-B1" country="EP" doc-number="1738753" kind="B1" lang="EN" date="20080423" family-id="37453347"
  date-produced="20100220" status="new">
  –<bibliographic-data>
    –<publication-reference fvid="88724218" ucid="EP-1738753-B1" status="new">
      –<document-id status="new" format="original">
          <country status="new">EP</country>
          <doc-number>1738753</doc-number>
          <kind>B1</kind>
          <date>20080423</date>
          <lang>EN</lang>
        </document-id>
      </publication-reference>
    +<application-reference mxw-id="PAPP77683688" ucid="EP-06017469-A" load-source="docdb" status="new" is-representative="NO"></application-
     reference>
    +<priority-claims status="new"></priority-claims>
    +<dates-of-public-availability status="new"></dates-of-public-availability>
    –<technical-data status="new">
      –<classifications-ipcr>
          <classification-ipcr mxw-id="PCL624787575" load-source="docdb" status="new">A61K 31/135 20060101C I20051008RMEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624787849" load-source="docdb" status="new">A61P 3/04 20060101ALI20051220RMJP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624795950" load-source="docdb" status="new">A61K 31/135 20060101A I20051008RMEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624799973" load-source="docdb" status="new">A61P 25/20 20060101ALI20051220RMJP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624806558" load-source="docdb" status="new">A61K 31/137 20060101CFI20071018BHEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624810330" load-source="docdb" status="new">A61K 31/137 20060101AFI20071018BHEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624820189" load-source="docdb" status="new">A61P 3/00 20060101CLI20051220RMJP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624827390" load-source="docdb" status="new">A61P 25/00 20060101ALI20071018BHEP </classification-ipcr>
          <classification-ipcr mxw-id="PCL624828549" load-source="docdb" status="new">A61P 25/00 20060101CLI20071018BHEP </classification-ipcr>
```

Baselines

A Patent document

Description, **claims**.



```
      <u style="single">Obesity Reduction Test Results</u>
    </b>
  </heading>
– <p num="p0023">
    The venlafaxine group showed consistent statistically significant mean weight decreases and mean percent decreases from baseline beginning at week 1.
    Overall, the mean decrease in body weight for the venlafaxine group at week 10 was 7.5 lb with a mean percent decrease from baseline of 3.6%. In
    contrast, the mean decrease in body weight for the placebo group at week 10 was 1.3 lb with a mean percent decrease from baseline of 0.7%. The body
    mass index evaluation for the venlafaxine also showed a pattern of decreases similar to that of the weight decreases.
  </p>
</description>
–<claims mxw-id="PCLM12825865" lang="DE" load-source="patent-office" status="new">
– <claim id="c-de-01-0001" num="0001">
  – <claim-text>
      Verwendung einer Verbindung mit der Formel
    + <chemistry id="chem0006" num="0006"></chemistry>
      in der A eine Komponente der Formel
    + <chemistry id="chem0007" num="0007"></chemistry>
      ist, wobei
      <br/>
      die gestrichelte Linie eine optionale Unsättigung darstellt;
    – <claim-text>
        R
        <sub>1</sub>
        Wasserstoff oder Alkyl mit 1 bis 6 Kohlenstoffatomen ist;
      </claim-text>
    – <claim-text>
        R
        <sub>2</sub>
```

Baselines

Parallel corpus selection

- Patent documents with **translated claims**.
  (not all of them!)
- IPC classification **A61P**.
  Specific therapeutic activity of chemical compounds or medical preparations.

Baselines

Parallel corpus selection

- Patent documents with **translated claims**.
  (not all of them!)
- IPC classification **A61P**.
  Specific therapeutic activity of chemical compounds or medical preparations.

56000 patents out of 1.3 million fulfill these demands.

(279282 aligned parallel fragments)

Baselines

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- The use according to claim 7, wherein said cancer diseases comprise bladder, lung, mamma, melanoma and prostate carcinomas.

- A compound according to claim 1 wherein it is (2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide.

- The pharmaceutical composition according to claim 1 or 2, wherein said platinum anticancer agent is selected from at least one of the complexes having structures of: **IMAGE**.

Baselines

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- **The use according to claim 7, wherein** said cancer diseases comprise bladder, lung, mamma, melanoma and prostate carcinomas.
- **A compound according to claim 1 wherein** it is (2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide.
- The pharmaceutical **composition according to claim 1 or 2, wherein said** platinum anticancer agent is selected from at least one of the complexes having structures of: **IMAGE**.

Baselines

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- The use according to claim 7, wherein said cancer diseases comprise **bladder, lung, mamma, melanoma and prostate carcinomas**.
- A compound according to claim 1 wherein it is (2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide.
- The pharmaceutical composition according to claim 1 or 2, wherein said **platinum anticancer agent** is selected from at least one of the complexes having structures of: **IMAGE**.

Baselines

Language domain and genre

Claims are written in a **lawyerish style** and using a very **specific vocabulary** of chemistry, full of **compounds names**.

Excerpt 1

- The use according to claim 7, wherein said cancer diseases comprise bladder, lung, mamma, melanoma and prostate carcinomas.

- A compound according to claim 1 wherein it is **(2S)-2-[(4S)-4-(2,2-difluorovinyl)-2-oxopyrrolidinyl]butanamide**.

- The pharmaceutical composition according to claim 1 or 2, wherein said platinum anticancer agent is selected from at least one of the complexes having structures of: **IMAGE**.

Baselines

Language domain and genre

Claims have also **long sentences** and **missing information**.

Excerpt 2

- Use of compounds of formula I **IMAGE** wherein R1 signifies substituted C1-C4-alkylene, whereby the substituents are selected from the group comprising unsubstituted aryloxy or aryloxy mono- to penta-substituted by R5, and unsubstituted pyridyloxy or pyridyloxy mono- to tetra-substituted by R5, whereby the substituents may be the same as one another or different if the number thereof is greater than 1; R2 signifies unsubstituted phenyl or phenyl mono- to penta-substituted by R5, or unsubstituted pyridyl or pyridyl mono- to tetra-substituted by R5; R3 is methyl; R4 signifies hydrogen, C1-C6-alkyl or halogen-C1-C6-alkyl; R5 signifies C1-C6-alkyl, C1-C6-alkoxy, halogen-C1-C6-alkyl, halogen-C1-C6-alkoxy, C2-C6-alkenyl, halogen-C2-C6-alkenyl, C2-C6-alkinyl, halogen-C2-C6-alkinyl, C3-C8-cycloalkyl, C1-C6-alkylcarbonyl, halogen-C1-C6-alkylcarbonyl, C1-C6-alkoxycarbonyl, halogen-C1-C6-alkoxycarbonyl, C1-C6-alkylsulfonyl, C1-C6-alkylsulfinyl, halogen, cyano or nitro; A signifies C(R6)(R7), CH=CH or C=C; R6 and R7 either, i ndependently of one another, signify hydrogen, halogen, C1-C6-alkyl, C1-C6-alkoxy, halogen-C1-C6-alkyl, halogen-C1-C6-alkoxy or C3-C6-cycloalkyl; or together signify C2-C6-alkylene; R8 and R9 are hydogen; m and n, independently...of one other, are 0 or 1; and optionally enantiomers thereof, with the proviso that if m is 0 then R1 is retained; in the preparation of a pharmaceutical composition for the control of endoparasitic helminths in warm-blooded productive livestock and domestic animals.

Baselines

SMT baseline, evaluation

BLEU

	EN2DE	DE2EN	EN2FR	FR2EN	DE2FR	FR2DE
Bing	0.33	0.43	0.43	0.45	0.20	0.24
Google	0.45	0.58	0.53	0.62	0.43	0.39
Domain	**0.58**	**0.65**	**0.62**	**0.70**	**0.56**	**0.53**

Baselines

English-German Translations, scores

	DE2EN			EN2DE		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1-WER	0.52	0.64	**0.72**	0.42	0.51	**0.69**
1-PER	0.66	0.76	**0.82**	0.56	0.64	**0.77**
1-TER	0.59	0.67	**0.76**	0.45	0.53	**0.71**
BLEU	0.43	0.58	**0.65**	0.33	0.45	**0.58**
NIST	8.25	9.67	**10.12**	6.53	8.05	**9.40**
ROUGE-W	0.40	0.48	**0.52**	0.34	0.41	**0.48**
GTM-2	0.30	0.40	**0.47**	0.25	0.32	**0.43**
METEOR-pa	0.60	0.69	**0.74**	0.36	0.45	**0.57**
ULC	0.09	0.29	**0.41**	0.03	0.19	**0.43**

Baselines

English-German Translations, examples

Why such good scores?

DE	Verwendung nach Anspruch 23 , worin das molare Verhältnis von Arginin zu Ibuprofen 0,60 : 1 beträgt .
EN	The use of claim 23 , wherein the molar ratio of arginine to ibuprofen is 0.60 : 1 .

Baselines

English-German Translations, examples

Why such good scores?

DE	Verwendung nach Anspruch 23 , worin das molare Verhältnis von Arginin zu Ibuprofen 0,60 : 1 beträgt .
EN	**The use** of claim 23 , wherein the molar ratio of arginine to ibuprofen **is** 0.60 : 1 .
Domain	The use of claim 23 , wherein the molar ratio of arginine to ibuprofen is 0.60 : 1 .
Google	The **method** of claim 23 , wherein the molar ratio of arginine to ibuprofen 0.60 : 1 **is** .
Bing	The Use of claim 23 , wherein the molar ratio of arginine to ibuprofen is 0.60 : 1 .

Baselines

English-German Translations, examples

What's wrong?

DE	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid
EN	(±)-N-(3-**a**minopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradeceneyloxy)-1-propanaminium **bromide**

Baselines

English-German Translations, examples

What's wrong?

DE	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid
EN	(±)-N-(3-aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradeceneyloxy)-1-propanaminium **bromide**
Domain	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid
Google	(±)-N-(3-aminopropyl)-N , N-dimethyl-2 , 3-bis (syn-9-tetradecenyloxy) is 1-propanaminiumbromid
Bing	(±)-N-(3-Aminopropyl)-N,N-dimethyl-2,3-bis(syn-9-tetradecenyloxy)-1-propanaminiumbromid

Baselines

English-French Translations, scores

	FR2EN			EN2FR		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1-WER	0.54	0.66	**0.78**	0.57	0.63	**0.73**
1-PER	0.71	0.78	**0.86**	0.68	0.75	**0.82**
1-TER	0.59	0.70	**0.80**	0.60	0.66	**0.74**
BLEU	0.45	0.62	**0.70**	0.43	0.53	**0.62**
NIST	8.52	10.01	**10.86**	8.39	9.21	**9.96**
ROUGE-W	0.41	0.50	**0.54**	0.39	0.45	**0.49**
GTM-2	0.32	0.43	**0.53**	0.31	0.36	**0.45**
METEOR-pa	0.61	0.72	**0.77**	0.57	0.65	**0.71**
ULC	0.07	0.28	**0.44**	0.10	0.23	**0.39**

Baselines

German-French Translations, scores

	DE2FR			FR2DE		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1-WER	0.42	0.52	**0.76**	0.30	0.43	**0.65**
1-PER	0.58	0.68	**0.77**	0.46	0.59	**0.74**
1-TER	0.47	0.56	**0.68**	0.32	0.46	**0.66**
BLEU	0.29	0.43	**0.56**	0.24	0.39	**0.53**
NIST	6.72	8.21	**9.10**	5.35	7.30	**8.88**
ROUGE-W	0.31	0.38	**0.45**	0.29	0.37	**0.44**
GTM-2	0.24	0.30	**0.41**	0.21	0.28	**0.41**
METEOR-pa	0.45	0.56	**0.64**	0.26	0.39	**0.51**
ULC	0.03	0.22	**0.41**	-0.03	0.19	**0.44**

Baselines

SMT Systems, general impressions (public systems)

Google

Few OOVs but tokenization problems with compounds.

Bing

Lack of specific vocabulary.

In-domain SMT

Try to solve the problems of the general systems, but still:

- Improve compound detector.
- Fix structures are translated different depending on the vocabulary.

Hybrid systems

Pros and Cons of the base systems

GF Pros (as compared to SMT)

- Capture **long distance** relations and reordering.
- Better **grammaticality.**

GF Cons (as compared to SMT)

- Dependence on the **initial parsing.**
- Lexical transfer **disambiguation.**
- High development **cost** of the grammars and associated resources.

Hybrid systems

Two hybridisation approaches

Statistical MT can alleviate some of the RBMT flaws

Hybrid systems

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Hybrid systems

Two hybridisation approaches

Rule-based MT can alleviate some of the SMT flaws

Who leads the hybrid model?

SMT. GF is used to enrich the "translation model" of the SMT system (known approach)

RBMT. SMT is used to provide confidence scored translation options to the RBMT target tree (novel)

Hybrid systems

A simple idea

Hard integration

Force fixed GF translations within a SMT system.

✓ Straightforward to implement from the SMT pov.

◇ Need of GF partial translations.

✗ There is no interaction between GF and SMT.

Hybrid systems

Another simple idea, hybrid SMT-GF system

SMT leads translation, GF complements

Complement the SMT translation table with GF options.

- If GF is able to generate Giza-like alignments, phrases can be extracted in the SMT way and we can combine translation tables.

GF alignments

- Based on the relation between the concrete syntaxes and the abstract syntax.
- Many-to-many.
- Semantic wrt. abstract syntax.

SMT alignments

- Based on corpus occurrences.
- One-to-many.

Hybrid systems

Alignment equivalence

From many-to-many to one-to-many

```
You want_to_go to the_nearest park
(0)     (1)    (2)     (3)    (4)

Quieres ir al parque mas cercano
(0)    (1)(2)  (3)   (4)   (5)

1-0 1-1 2-2 3-4 3-5 4-3
```

(alignments from Phrasebook grammar)

Conclusions

Summary

- The first step towards hibridisation has been building individual systems.
- SMT already achieves an acceptable translation quality.
- However, the combination of different approaches to translation can help to solve the observed translation errors.
- Several ways to combine GF and SMT can (and should!) be applied.

SMT within MOLTO's hybrid translation system

Cristina España-Bonet

Universitat Politècnica de Catalunya, TALP Research Center

–GF Summer School–

Barcelona, August 25th, 2011

Hybrid SMT-RBMT: Experiments

Phrasebook grammar (toy example)

- Syntetic corpus generation.
- Parallel corpus with 200 sentences.
- Insignificant for SMT (by 2-3 orders of magnitude!).
- Null intersection with SMT corpora.

Patents grammar

- Needed for real experiments.

Conclusions

Hybrid SMT-RBMT: Experiments

Translation Table, core of an SMT system:

```
source language ||| target language ||| probabilities

...
quite a burden ||| un estorbo muy grande ||| 0.25 1.57587e-06 0.25 3.57895e-12 2.718
quite a burden ||| un estorbo muy ||| 0.25 1.57587e-06 0.25 8.38161e-08 2.718
quite a challenge but we ||| todo un reto , pero lo ||| 0.5 6.64558e-05 1 1.46764e-06 2.718
quite a challenge but ||| todo un reto , pero ||| 0.5 0.00179307 1 9.70607e-05 2.718
quite a challenge ||| todo un reto , ||| 0.5 0.002396 0.5 0.000190619 2.718
quite a challenge ||| todo un reto ||| 0.333333 0.002396 0.5 0.00244338 2.718
quite a considerable delay ||| un retraso muy considerable ||| 0.333333 2.91692e-05 ...
quite a contribution towards ||| una importante contribución en lo ||| 0.25 9.69758e-07 ...
quite a contribution towards ||| una importante contribución en ||| 0.142857 9.69758e-07 ...
quite a difference whether ||| muy diferente ||| 0.0344828 8.29695e-09 1 0.0013126 2.718
quite a difference ||| muy diferente ||| 0.0344828 1.38144e-05 1 0.0013126 2.718
...
```

Conclusions

Hybrid SMT-RBMT: Experiments on combination

GF scored partial output as **new features** in SMT decoding.

$$\log P(e|f) \sim \lambda_{lm} \log P(e) + \lambda_g \log P(f|e) + \lambda_d \log P(e|f)$$
$$+\lambda_{di} \log P_{di}(e,f) + \lambda_w \log w(e) + \lambda_{\mathbf{GF}} \mathbf{log\, P_{GF}(e|f)}$$

`quite a challenge|||todo un reto|||0.333 0.002 0.5 0.002 2.718` $\mathbf{log\, P}_{\mathrm{GF}}(e|f)$

Requirements:

- GF predictions have to be probabilistic.
- Phrase pairs without prediction must be complemented.

RBMT leads translation, SMT decodes

Complement the RBMT translation structure with SMT options.

- **SMatxinT**

  Approach being applied for **Basque-to-Spanish** with the RBMT system Matxin.

  `OpenMT-2 Spanish Research Project`
  `UPC+EHU collaboration`

Conclusions

An hybrid RBMT-SMT system: SMatxinT, methodology

- The RBMT system must parse and translate the input sentence.
- Phrases and segmentation are those given by the RBMT system.
- Each segment (and up) is sent to a generic SMT to provide more partial translations.
- A Moses-like decoder is fed with the resulting phrases to search for the highest scored translation.
- This statistical decoder performs no reordering and uses very simple features.

Conclusions

An hybrid RBMT-SMT system: SMatxinT, comments

Current results

- Large difference between in-domain and out-of-domain scenarios.
- Results are at most close to SMT system.
- Oracles show large room for improvement.
- RBMT phrases are underused.
- Current features are not distinctive enough.

Conclusions

SMatxinT in relation with MOLTO

SMatxinT vs. MOLTO

General translator vs. in-domain translator

With SMatxinT results are better for out-of-domain tests, where the difference between SMT and RBMT systems is less important, but systems (specially SMT) have a lower quallity.

Matxin vs. GF

General grammar vs. in-domain grammar

With MOLTO both systems will be in-domain, so they are expected to be high quality. Improvements here will be over already good translations.

Conclusions

Statistical extension of GF grammar

Learning GF grammars

Abstract syntax	Like She He	Grammarian
English example	she likes him	Grammarian
German translation	er gefällt ihr	**SMT**
Resource tree	mkCl he$_{\mathrm{Pron}}$ gefallen$_{\mathrm{V2}}$ she$_{\mathrm{Pron}}$	GF parser
Syntax rule	Like x y $=$ mkCl y gefallen$_{\mathrm{V2}}$ x	Variables renamed

- SMT of short and frequent sentences is good

Conclusions

Statistical extension of GF grammar, application

- Applied to the **Phrasebook grammar**
- **Languages**: Danish, Dutch, German, Norwegian
- Phrasebook **demo**:
  `http://www.molto-project.eu/demo/phrasebook`