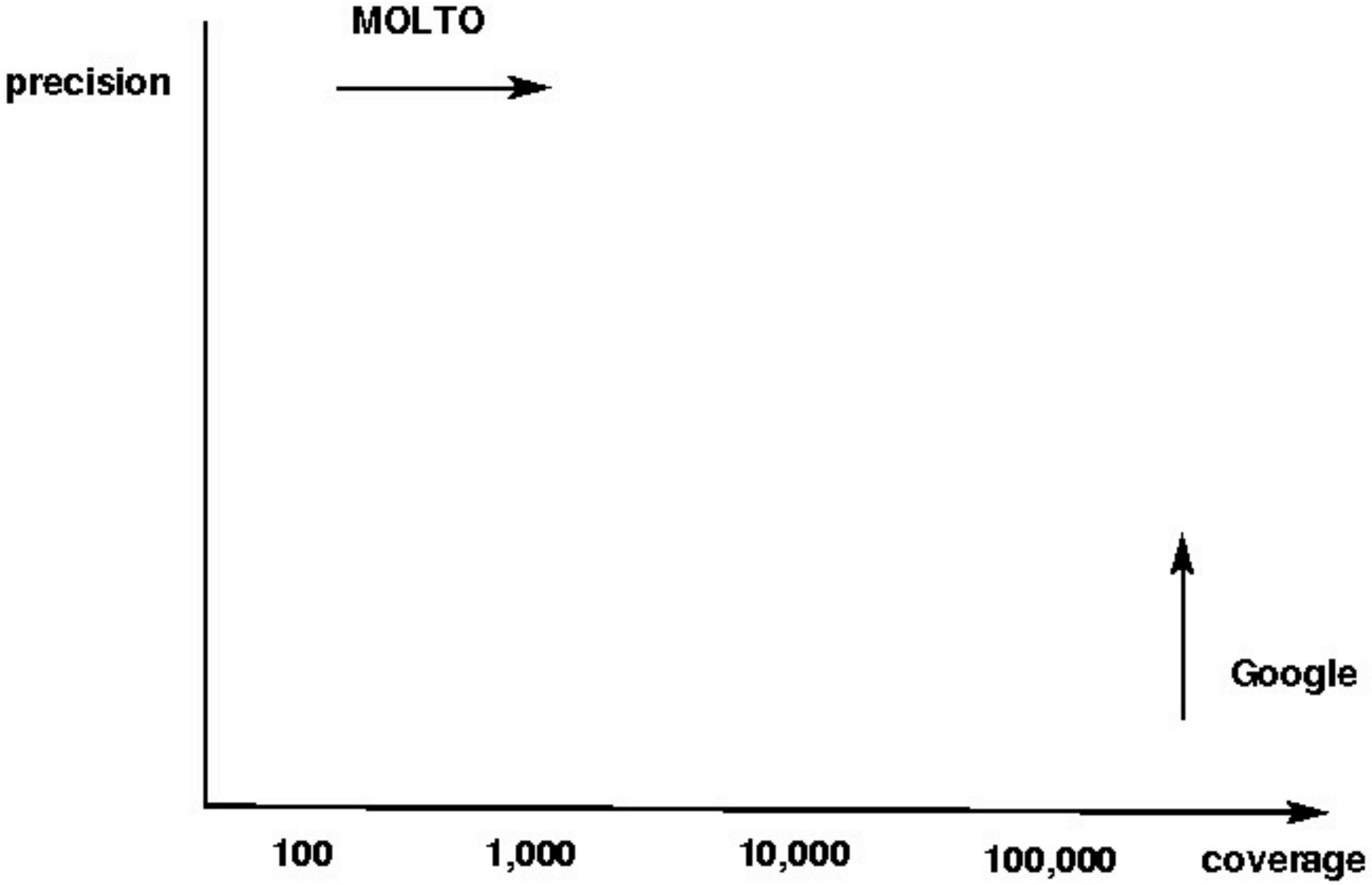
What can GF learn from SMT

Ramona Enache
University of Gothenburg

MOLTO



GF/SMT

- GF

  - high-quality

  - limited coverage

- SMT

  - variable quality

  - large coverage

GF

- Advantages
  - many languages in parallel
  - underesourced languages
  - easy to fix
  - generalizable

GF

- Disadvantages

  - development is labour intensive

  - need language skills + programming skills

  - limited coverage

  - literal translations (with resource grammars)

GF

- Needs improvements on
  - coverage
    - syntactic structures
    - lexical items
  - development effort for
    - abstract syntax
    - concrete syntax

SMT

- Advantages

  - less human effort for development

  - large coverage domain-wise

  - robust towards ungrammatical input

  - good for idiomatic and common expressions

SMT

- Disadvantages

  - long distance dependencies

  - underesourced languages/sparse training data

  - bilingual

  - could produce ungrammatical output

GF + SMT 1

- Direction 1

  - developing a domain-specific concrete syntax with input from a SMT system/informed user

  - efficient for short idiomatic formulations

GF + SMT 1

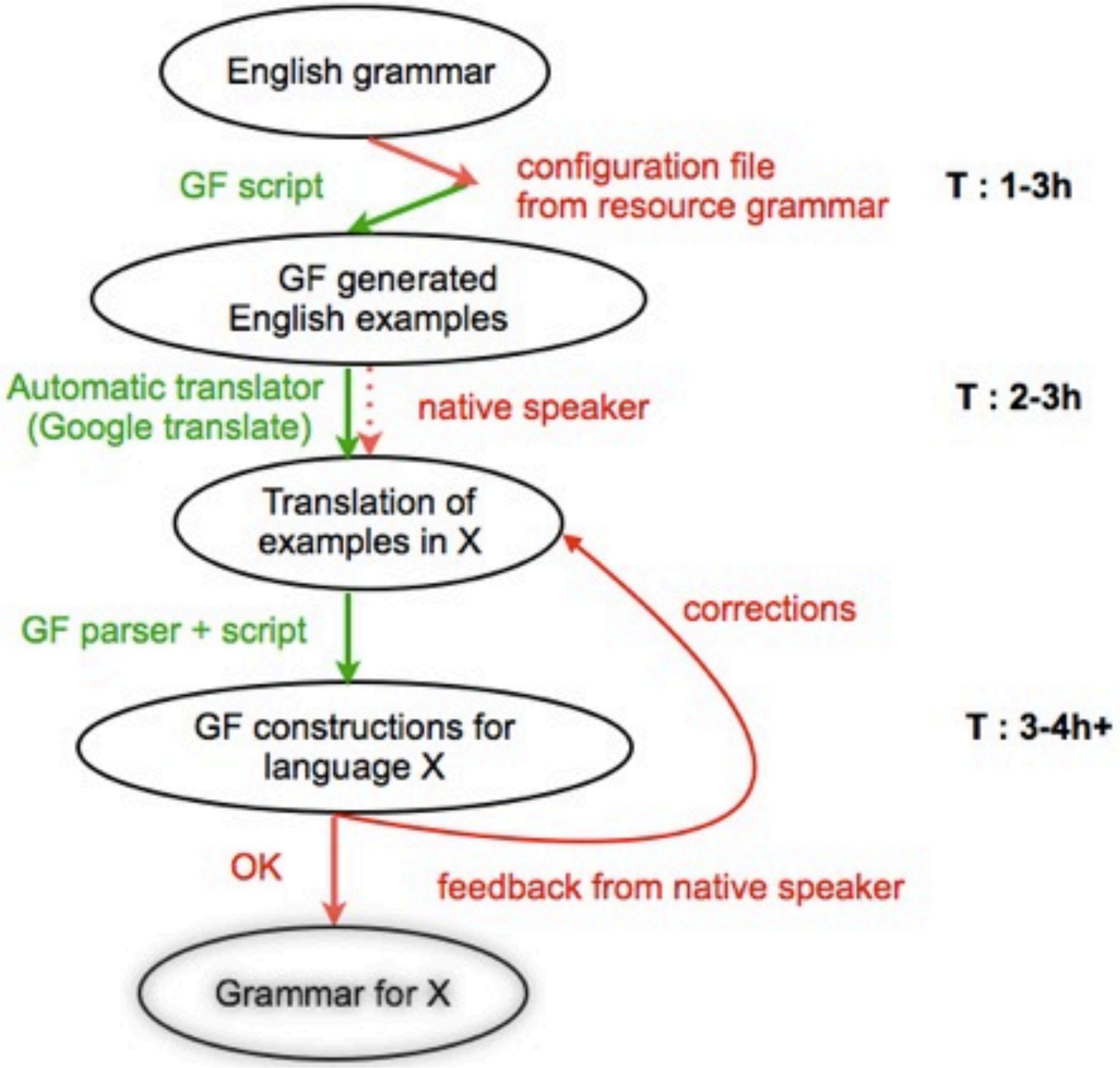
- Use case

  - MOLTO Phrasebook

  - 5 languages (German, Danish, Dutch, Norwegian, Polish)

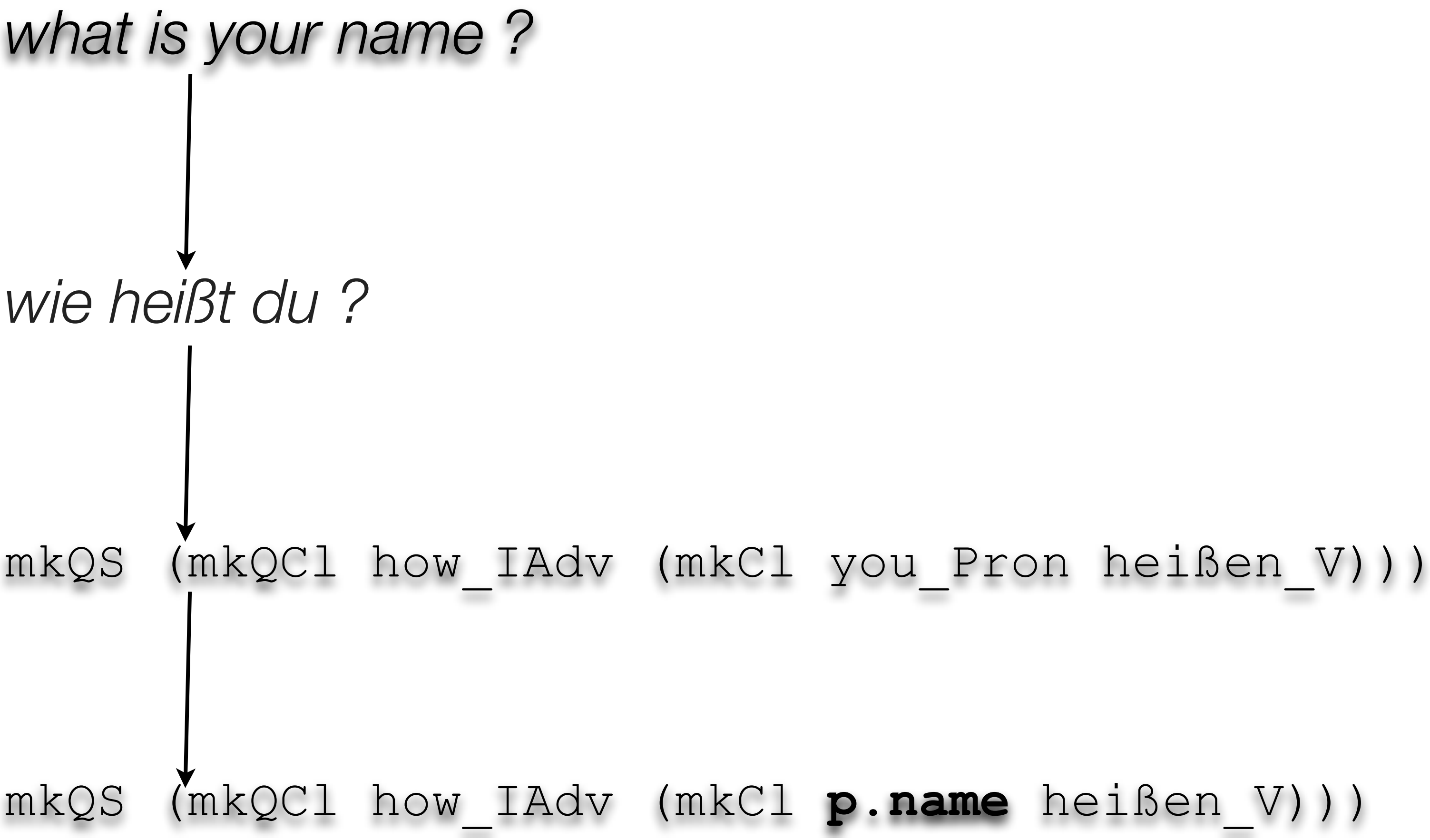
  - has English Phrasebook as starting point

GF + SMT 1


English grammar
GF script
configuration file
from resource grammar
T : 1-3h
GF generated
English examples
Automatic translator
(Google translate)
native speaker
T : 2-3h
Translation of
examples in X
corrections
GF parser + script
GF constructions for
language X
T : 3-4h+
OK
feedback from native speaker
Grammar for X

GF + SMT 1

- Example (function that models asking about people’s name in German)



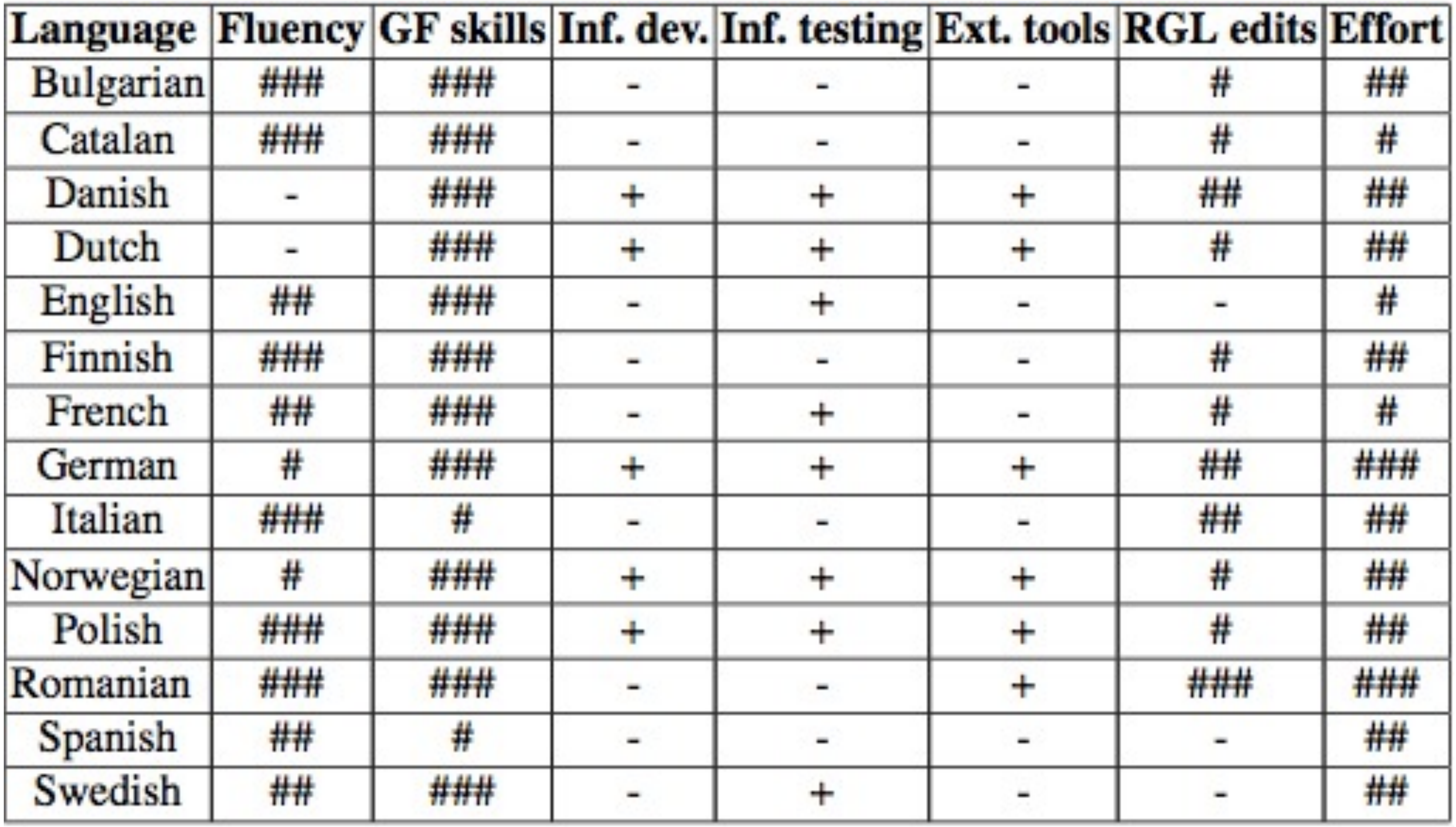
what is your name ?

wie heißt du ?

```
mkQS (mkQCl how_IAdv (mkCl you_Pron heißen_V)))
```

```
mkQS (mkQCl how_IAdv (mkCl p.name heißen_V)))
```

GF + SMT 1



Language	Fluency	GF skills	Inf. dev.	Inf. testing	Ext. tools	RGL edits	Effort
Bulgarian	###	###	-	-	-	#	##
Catalan	###	###	-	-	-	#	#
Danish	-	###	+	+	+	##	##
Dutch	-	###	+	+	+	#	##
English	##	###	-	+	-	-	#
Finnish	###	###	-	-	-	#	##
French	##	###	-	+	-	#	#
German	#	###	+	+	+	##	###
Italian	###	#	-	-	-	##	##
Norwegian	#	###	+	+	+	#	##
Polish	###	###	+	+	+	#	##
Romanian	###	###	-	-	+	###	###
Spanish	##	#	-	-	-	-	##
Swedish	##	###	-	+	-	-	##

GF + SMT 2

- Direction 2

  - using SMT lexical tables to build a bilingual lexicon for GF grammars

  - for main lexical categories N, A, Adv, (V ?)

  - aims to improve grammar coverage in terms of lexical items

GF + SMT 2

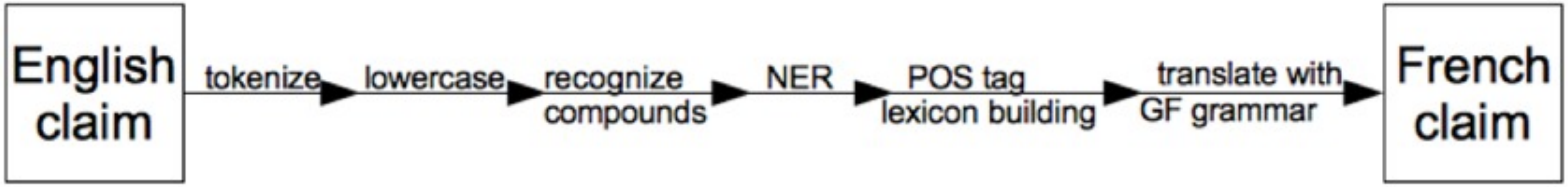
- Use case

  - translation of patent claims from the biomedical domain from English to French and German

  - many unknown words to the monolingual dictionaries

  - need large specialized bilingual dictionaries

GF + SMT 2


English claim
tokenize
lowercase
recognize compounds
NER
POS tag
lexicon building
translate with GF grammar
French claim

GF + SMT 2

- Result

  - dictionary can be created at runtime, based on the corpus to translate or before, based on the whole training corpus

  - good quality, but for publishing quality, it needs human check and post-editing

  - no clear solution for V2, especially with prepositions

GF + SMT 2

First approach **Runtime-Safe**

- use word pairs where each component can be analyzed by the monolingual lexicon

- starts from a small hand-crafted lexicon with the most frequent words in the training corpus

- one-to-one translations

GF + SMT 2

Second approach **Runtime-Unsafe**

- use word pairs regardless if they are found or not in the monolingual lexica

- starts from the same core lexicon

- one-to-one translations

GF + SMT 2

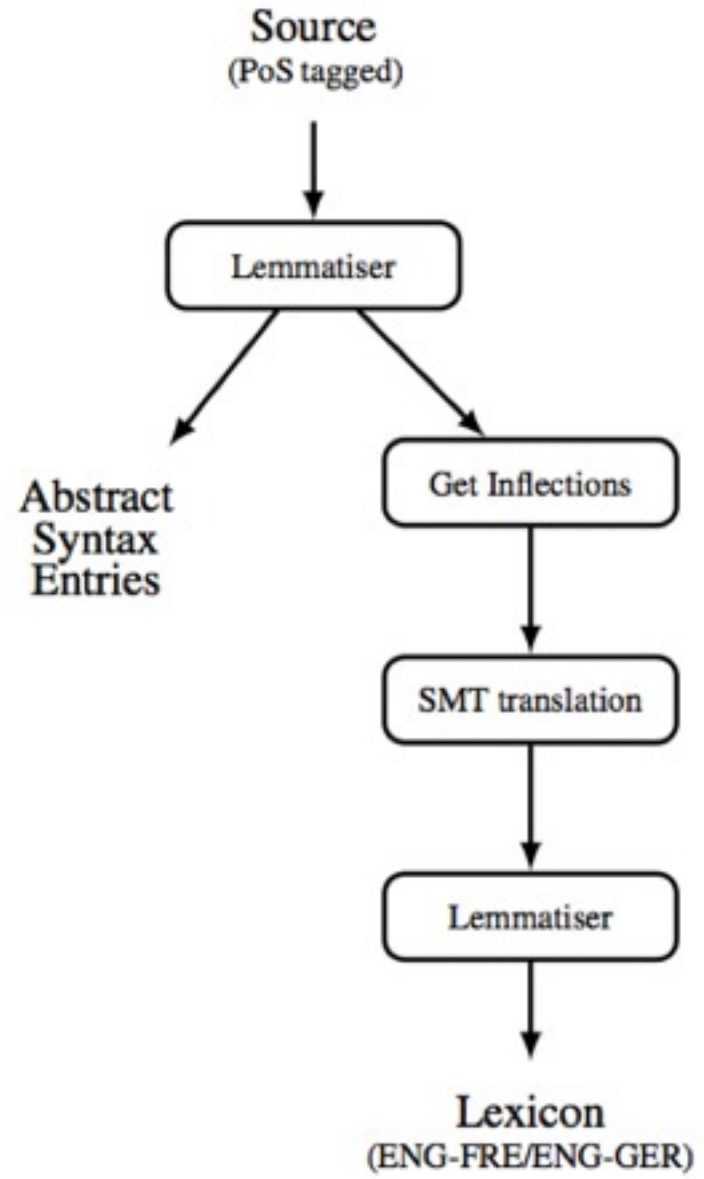
Third approach **Static**

- English-French built completely from the translation tables

- 3,983 entries (N, A, Adv)

- one-to-many translations

GF + SMT 2


Source
(PoS tagged)
Lemmatiser
Abstract Syntax Entries
Get Inflections
SMT translation
Lemmatiser
Lexicon
(ENG-FRE/ENG-GER)

GF + SMT 2

- Example:

  *A human monoclonal **antibody** according to any of the claims 1-6, characterized in that the antibody is an IgG1 molecule.*

GF + SMT 2

- Example:

  Genia: ***antibody*** -> *N*

  - We add `antibody_N : N;` to abstract syntax

  - We find `antibody_N` in DictEng.gf and infer *antibodies, antibody's, antibodies'*

GF + SMT 2

- Example:

  - Lexical tables :

  `anticorps antibody 0.4774548`

  - We find `anticorps_N` in DictFre.gf and infer the gender and the plural form (*anticorps*)

  - We add the pair `(DictEng.antibody_N, DictFre.anticorps_N)` to the lexicon.

GF + SMT 3

- Direction 3

  - extracting multiword expressions from SMT translation tables

  - use case - German compound noun phrases

GF + SMT 3

- Approach

 - create grammar with rules for compounding in German

+ w1 + lowercase(w2)

+ w1 + ‘s’ + lowercase(w2)

+ w1 + ‘-’ + w2

+ ...

GF + SMT 3

- Approach

  - find compound candidates in translation table

    + 1 word in German, more than 1 words in English translation

    + confidence score > threshold

    + English translation is parseable as CN with GF resource grammar(no robustness)

GF + SMT 3

- Approach (cont'd)

  - German compound is split into components with a greedy approach

    + if the word is in DictGer.gf - stop

    + owise, backtrack to find the smallest number of splits which yield the largest words which belong to DictEng.gf and the combination is parsed by the compounds grammar defined before

GF + SMT 3

- Approach (cont'd)

  - add compound parsed with compound grammar + English phrase parsed as CN

  - abstract syntax - from English

  - one-to-many

GF + SMT 3

- Approach (cont’d)
    - 7,774 lexical items added in this manner
    - not fully evaluated - help please :-)

GF + SMT 3

- Example

  Bauchchirurgie -> *Bauch* + *Chirurgie*

  English: abdominal surgery

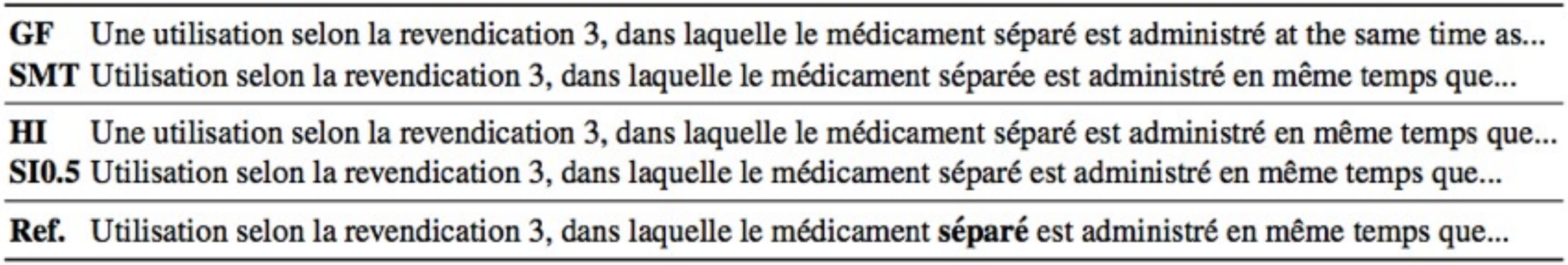
GF + SMT 4

- GF + SMT hybrid system for patent translation

GF + SMT 4

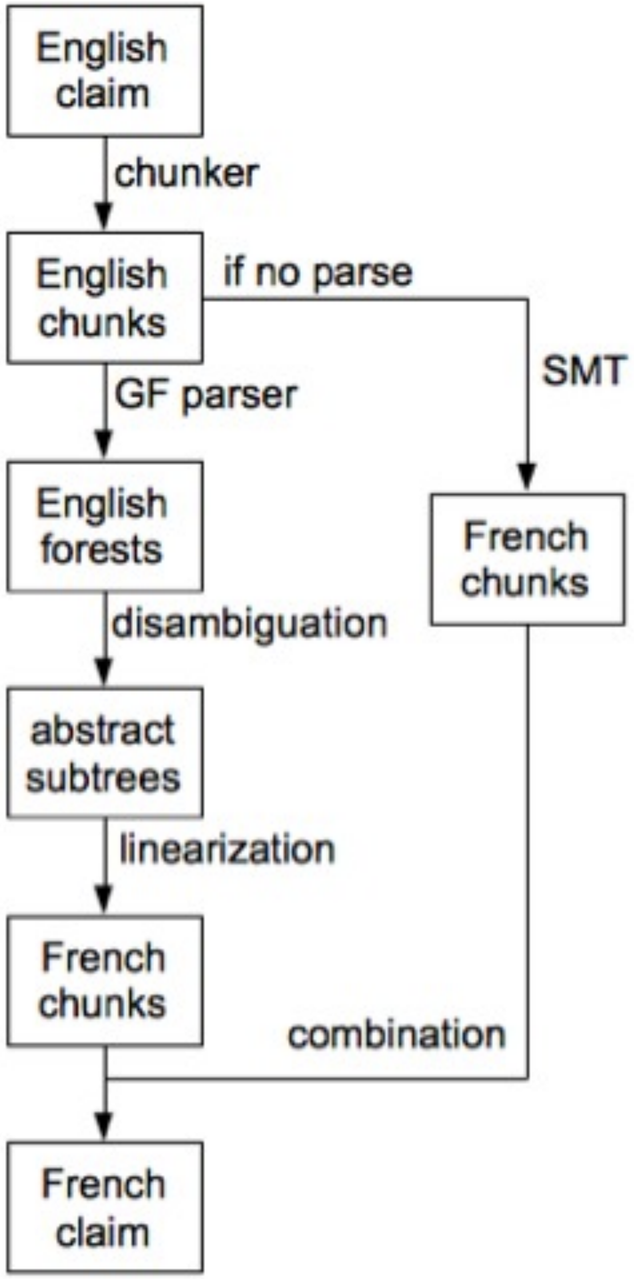
- aim to get the large coverage of SMT and high-precision of GF
- in particular for correcting syntax errors that could affect understanding (correct agreements)

GF + SMT 4



GF	Une utilisation selon la revendication 3, dans laquelle le médicament séparé est administré at the same time as...
SMT	Utilisation selon la revendication 3, dans laquelle le médicament séparée est administré en même temps que...
HI	Une utilisation selon la revendication 3, dans laquelle le médicament séparé est administré en même temps que...
SI0.5	Utilisation selon la revendication 3, dans laquelle le médicament séparé est administré en même temps que...
Ref.	Utilisation selon la revendication 3, dans laquelle le médicament **séparé** est administré en même temps que...

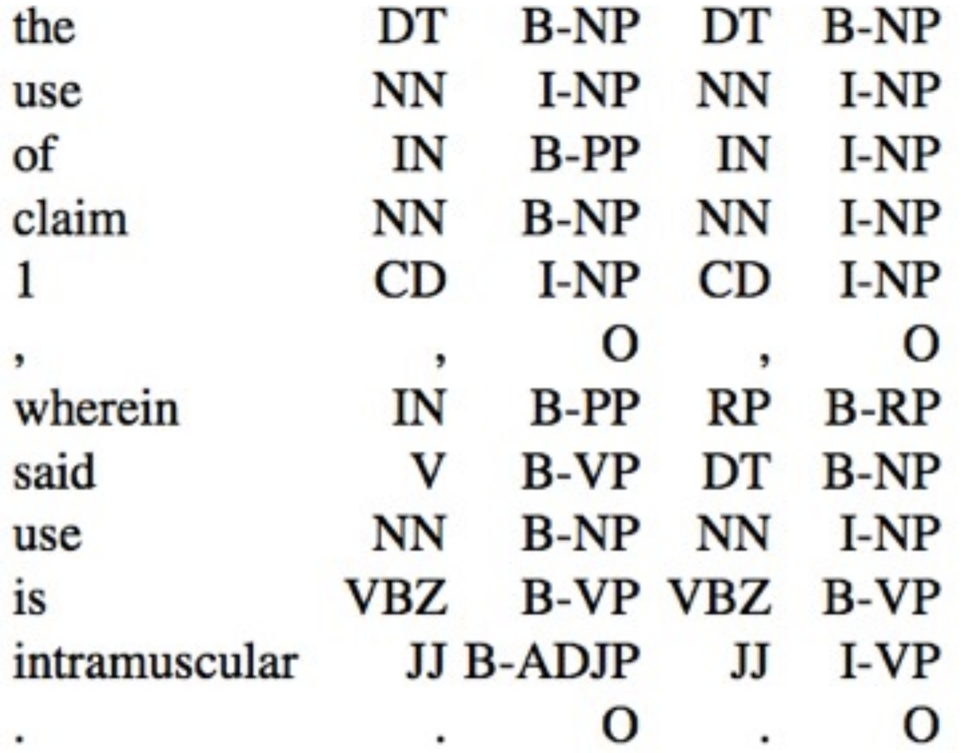
GF + SMT 4


English claim
chunker
English chunks
if no parse
SMT
GF parser
English forests
French chunks
disambiguation
abstract subtrees
linearization
French chunks
combination
French claim

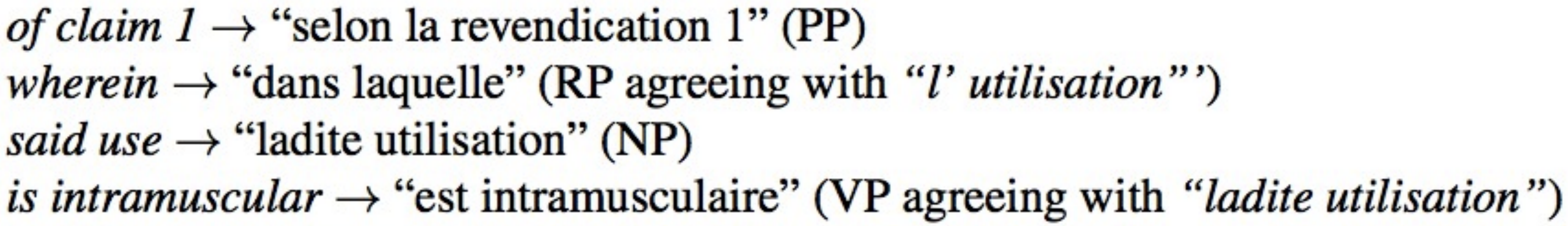
GF + SMT 4

Example:

the use of claim 1, wherein said use is intramuscular.



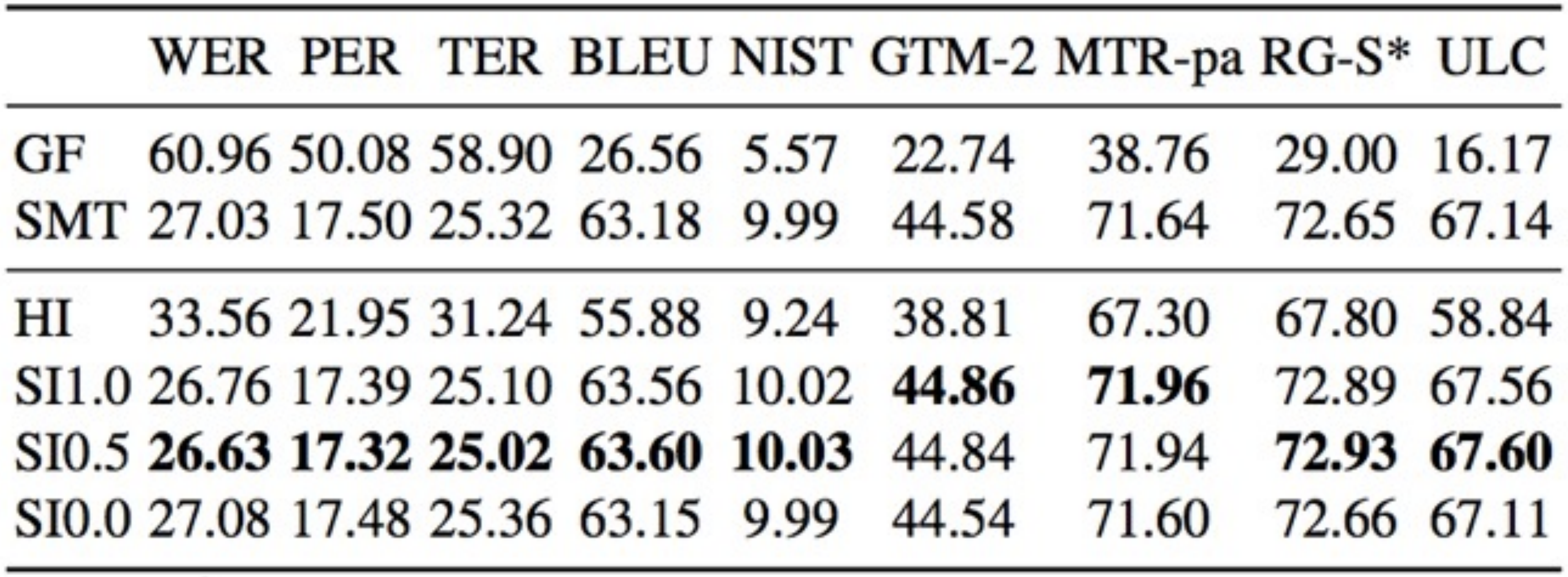
the	DT	B-NP	DT	B-NP
use	NN	I-NP	NN	I-NP
of	IN	B-PP	IN	I-NP
claim	NN	B-NP	NN	I-NP
1	CD	I-NP	CD	I-NP
,	,	O	,	O
wherein	IN	B-PP	RP	B-RP
said	V	B-VP	DT	B-NP
use	NN	B-NP	NN	I-NP
is	VBZ	B-VP	VBZ	B-VP
intramuscular	JJ	B-ADJP	JJ	I-VP
.	.	O	.	O



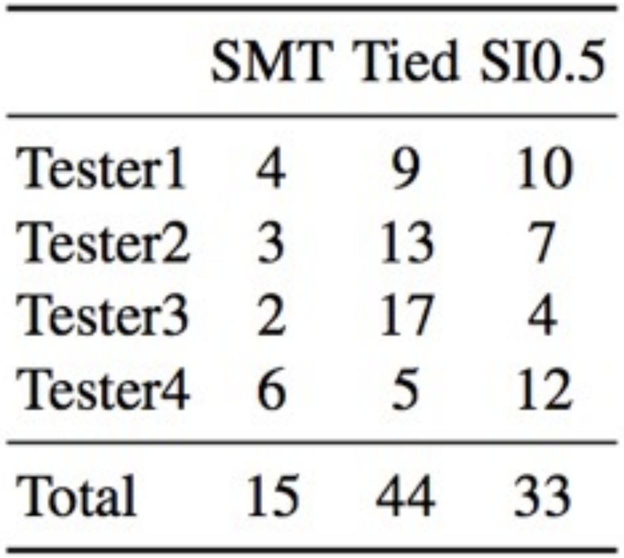
the use → "l' utilisation" (NP)
of claim 1 → "selon la revendication 1" (PP)
wherein → "dans laquelle" (RP agreeing with *"l' utilisation"*)
said use → "ladite utilisation" (NP)
is intramuscular → "est intramusculaire" (VP agreeing with *"ladite utilisation"*)

GF + SMT 4

Evaluation(English-French)



	WER	PER	TER	BLEU	NIST	GTM-2	MTR-pa	RG-S*	ULC
GF	60.96	50.08	58.90	26.56	5.57	22.74	38.76	29.00	16.17
SMT	27.03	17.50	25.32	63.18	9.99	44.58	71.64	72.65	67.14
HI	33.56	21.95	31.24	55.88	9.24	38.81	67.30	67.80	58.84
SI1.0	26.76	17.39	25.10	63.56	10.02	**44.86**	**71.96**	72.89	67.56
SI0.5	**26.63**	**17.32**	**25.02**	**63.60**	**10.03**	44.84	71.94	**72.93**	**67.60**
SI0.0	27.08	17.48	25.36	63.15	9.99	44.54	71.60	72.66	67.11



	SMT	Tied	SI0.5
Tester1	4	9	10
Tester2	3	13	7
Tester3	2	17	4
Tester4	6	5	12
Total	15	44	33

Conclusion

- GF grammar development can be enhanced by using SMT tools
- Lexicon acquisition and concrete grammar building give promising results
- hybrid translation systems do not improve over SMT for the given language pair and domain, but more GF-SMT combinations are possible.