WP5
Statistical and Robust Translation

Cristina España-Bonet

Universitat Politècnica de Catalunya, TALP Research Center

– 5th project meeting –

Utrecht, September 20th, 2012

WP5

Overview

1. General view
2. Ongoing work
3. Future work
4. Dissemination

General view

Goal

Extension of the grammar-based translation methods to widen their coverage and quality in unconstrained text translation.

General view

Goal

Extension of the grammar-based translation methods to widen their coverage and quality in unconstrained text translation.

Especially related to:

WP2 Grammar-based translation method

WP7 Quasi-unconstrained domain, patents

WP9 Evaluation

General view

Participants & PMs & Tasks

UPC	38

SMT technology, hybrid models, corpora processing, evaluation

General view

Participants & PMs & Tasks

UPC **38**	SMT technology, hybrid models, corpora processing, evaluation
UGOT **9**	Probabilistic extension of GF, synthetic corpora for SMT

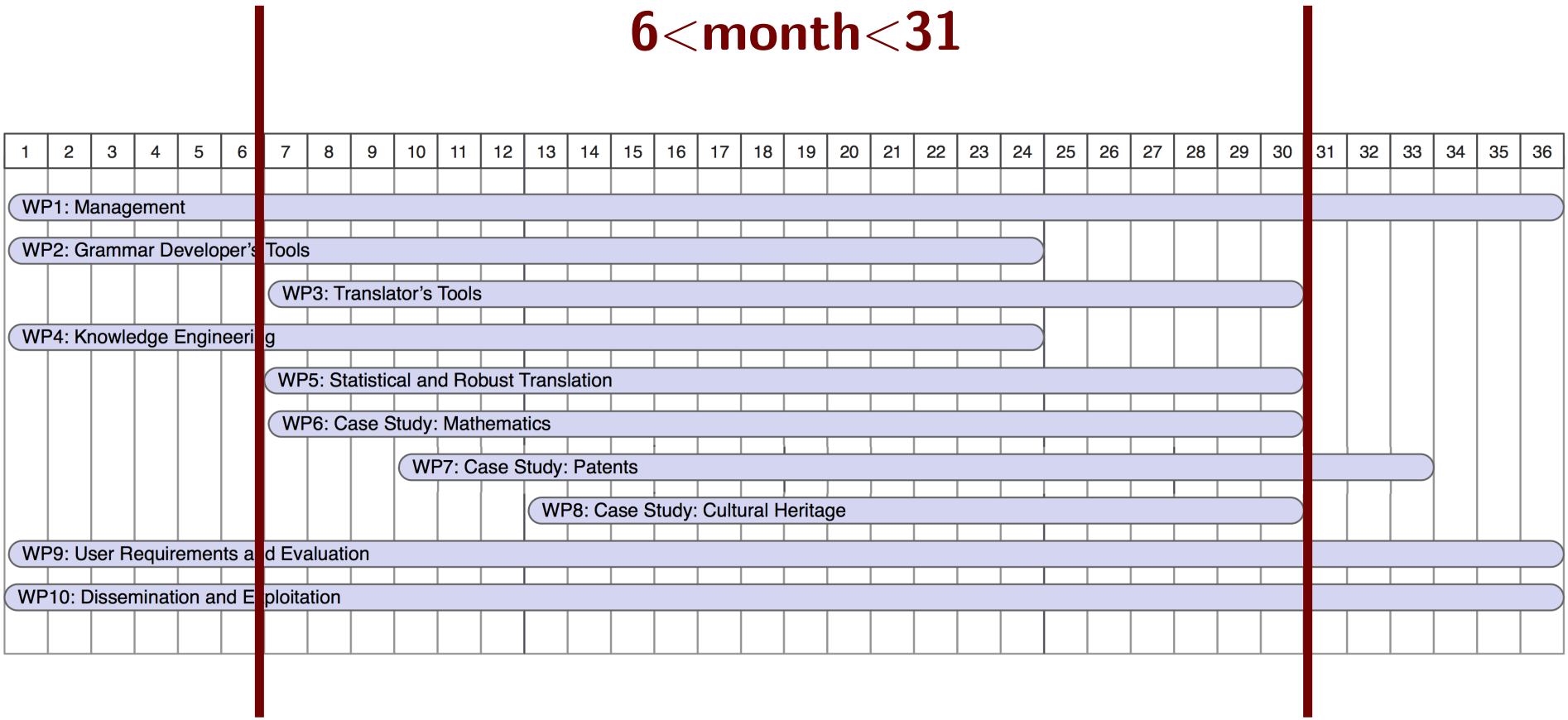
General view

Participants & PMs & Tasks

UPC	**38**	SMT technology, hybrid models, corpora processing, evaluation
UGOT	**9**	Probabilistic extension of GF, synthetic corpora for SMT
UHEL	**6**	Usability and evaluation of the combined system

General view

Timeline



General view

Milestones & Deliverables

Month 18 — Month 24 — Month 30

MS5

First prototypes of the *baseline* combination models.

D51

Description of the final collection of corpora.

General view

Milestones & Deliverables

Month 18 — **Month 24** — Month 30

MS7

First prototypes of hybrid combination models.

D52

Description and evaluation of the combination prototypes.

General view

Milestones & Deliverables

Month 18 — Month 24 — **Month 30**

MS8

Translation tool complete.

D53

WP5 final report: statistical and robust MT.

Ongoing work

Overview

Ongoing work

From M24 to M30, highlights

- Improve French grammar
- Build German GF & Hybrid system
- Extend robust parsing
- Develop new hybrids
- One-click system

Ongoing work

Improving the grammar (French)

Version 1: original grammar, manual intervention on the lexicon, and Genia tokeniser

Ongoing work

Improving the grammar (French)

Version 1: original grammar, manual intervention on the lexicon, and Genia tokeniser

Version 2: modifications to the grammar, reduced size of the lexicon, in-house tokeniser

Version 3: modifications to the grammar, full lexicon with associated probabilities, in-house tokeniser

Ongoing work

Improving the grammar (French)

	Version 1	Version 2	Version 3
WER	61.16	63.99	**60.32**
PER	51.34	53.78	**47.62**
TER	59.92	62.54	**58.59**
BLEU	**26.47**	23.34	26.22
NIST	5.55	5.14	**5.65**
GTM-2	**22.54**	20.51	21.97
MTR-ex	39.84	37.36	**41.77**
MTR-pa	38.80	36.61	**40.99**
RG-S*	29.00	24.75	**30.46**
ULC	16.75	12.45	**18.61**

Ongoing work

German GF grammar & homogenisation

- German grammar built from scratch
- Uses standard RGL German
- French Grammar modified accordingly

Ongoing work

German GF grammar & homogenisation

- German grammar built from scratch
- Uses standard RGL German
- French Grammar modified accordingly

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-pa	RG-S*	ULC
GF-vX	71.39	60.76	69.78	23.53	5.13	20.80	42.18	26.05	17.33
SMT	30.93	22.82	29.33	57.59	9.40	42.98	57.08	63.14	63.90

Ongoing work

Development of more evolved hybrid systems

	Integration	**Grammar**		**GF confidence**			**GF on dev**
			1 level	low	med	high	
HI-v1	hard	V1	○	○	○	○	○
SI1.0-v1	soft	V1	1.0	○	○	○	○
SI0.5-v1	soft	V1	0.5	○	○	○	○
SI0.0-v1	soft	V1	0.0	○	○	○	○

Ongoing work

Development of more evolved hybrid systems

	Integration	Grammar	GF confidence				GF on dev
			1 level	low	med	high	
HI-v1	hard	V1	○	○	○	○	○
SI1.0-v1	soft	V1	1.0	○	○	○	○
SI0.5-v1	soft	V1	0.5	○	○	○	○
SI0.0-v1	soft	V1	0.0	○	○	○	○
HI-v3	hard	V3	○	○	○	○	○
SIp1-v3	soft	V3	○	0.2	1	4	○
SIp2-v3	soft	V3	○	0.2	5	50	○
SIp3-v3	soft	V3	○	0.2	50	500	○
SIp4-v3	soft	V3	○	0.2	5000	50000	○
SIp5-v3	soft	V3	○	0.2	500000	5000000	○
HI-v3d	hard	V3	○	○	○	○	●
SIp1-v3d	soft	V3	○	0.2	1	4	●
SIp3-v3d	soft	V3	○	0.2	50	500	●
SIp5-v3d	soft	V3	○	0.2	500000	5000000	●

Ongoing work

Development of more evolved hybrid systems

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-pa	RG-S*	ULC
GF-v1	61.16	51.34	59.92	26.47	5.55	22.54	38.80	29.00	15.74
GF-v3	60.32	47.62	58.59	26.22	5.64	21.97	40.99	30.46	17.63
SMT	27.25	18.28	25.69	62.30	9.94	44.90	71.59	72.65	66.32
HI-v1	33.17	22.47	31.23	55.37	9.26	39.14	67.05	67.80	58.55
SI1.0-v1	26.93	18.20	25.43	62.69	9.98	**45.24**	71.82	72.89	66.77
SI0.5-v1	**26.78**	**18.12**	**25.33**	**62.75**	**9.99**	45.22	**71.83**	**72.93**	**66.82**
SI0.0-v1	27.26	18.30	25.70	62.27	9.94	44.89	71.55	72.66	66.32
HI-v3	40.70	29.03	38.65	45.97	8.13	33.15	60.41	56.48	46.72
SIp1-v3	27.22	18.58	25.74	62.21	9.91	44.80	71.62	72.06	66.12
SIp2-v3	27.61	18.88	26.12	61.71	9.86	44.09	71.32	71.63	65.45
SIp3-v3	28.00	19.17	26.51	61.21	9.82	43.56	70.94	71.10	64.80
SIp4-v3	28.43	19.50	26.92	60.54	9.75	43.10	70.55	70.23	64.00
SIp5-v3	29.00	19.88	27.43	59.87	9.67	42.32	70.11	69.74	63.16
HI-v3d	40.47	28.61	38.39	46.29	8.16	33.41	60.47	56.65	47.07
SIp1-v3d	26.97	18.21	25.50	62.37	9.95	44.53	71.57	72.68	66.37
SIp3-v3d	27.49	18.56	25.98	61.64	9.89	43.92	71.12	71.99	65.53
SIp5-v3d	28.83	19.46	27.25	60.04	9.71	42.57	70.06	70.53	63.61

Ongoing work

Development of more evolved hybrid systems

English-to-French

- Hybrid outperforms SMT & GF, but
- improvements on the system are not reflected into the translations

Ongoing work

Development of more evolved hybrid systems

English-to-French

- Hybrid outperforms SMT & GF, but
- improvements on the system are not reflected into the translations

English-to-German

- Hybrid at the same level as the SMT

Ongoing work

Development of more evolved hybrid systems

English-to-German

- Hybrid at the same level as the SMT

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-pa	RG-S*	ULC
GF-vX	71.39	60.76	69.78	23.53	5.13	20.80	42.18	26.05	17.33
SMT	**30.93**	**22.82**	**29.33**	**57.59**	**9.40**	**42.98**	57.08	63.14	63.90
SI0.5-v1	31.42	23.09	29.84	57.36	9.38	42.96	**67.40**	**62.63**	**66.36**

Ongoing work

Development of more evolved hybrid systems

English-to-German

- Hybrid at the same level as the SMT

	WER	PER	TER	BLEU	NIST	GTM-2	MTR-pa	RG-S*	ULC
GF-vX	71.39	60.76	69.78	23.53	5.13	20.80	42.18	26.05	17.33
SMT	**30.93**	**22.82**	**29.33**	**57.59**	**9.40**	**42.98**	57.08	63.14	63.90
SI0.5-v1	31.42	23.09	29.84	57.36	9.38	42.96	**67.40**	**62.63**	**66.36**

- Frequent errors in chunking (Genia) $\Rightarrow$ **Robust Parsing**

Ongoing work

One-click system

Available at UGOT's server

```
csmisc14:hybrid cristina$ perl H1PTrad.pl

Usage: perl H1PTrad.pl -v # <input> [src2trg]
       -v: verbosity [0,1,2]
       input: file to translate
       src2trg: language pair

Ex: perl H1PTrad.pl -v 1 /systems/hybrid/input/patsA61P.test.en en2fr
```

(Demo afterwards)

Ongoing work

One-click system

Run time (s)

	1 segment	10 segments	100 segments	1000 segments
Tokenisation	0	0	3	25
GF translation	219	209	224	360
Filtering	33	33	35	40
SMT translation	10	22	217	2549
Total	263	266	480	2975

Ongoing work

Flagship

‘‘A hybrid patent translation system that beats its competitors, if possible’’

Ongoing work

Flagship

‘‘A hybrid patent translation system that beats its competitors, if possible’’

Comparison with Bing, Google, PLuTO

Not as easy as it sounds:

- We share corpus (therefore problems with test sets)

Future work

Overview

Include Robust Parsing

- Should be better and better integrated than an external software like Genia

Future work

Hybrid translator: Robustness by GF

Include Robust Parsing

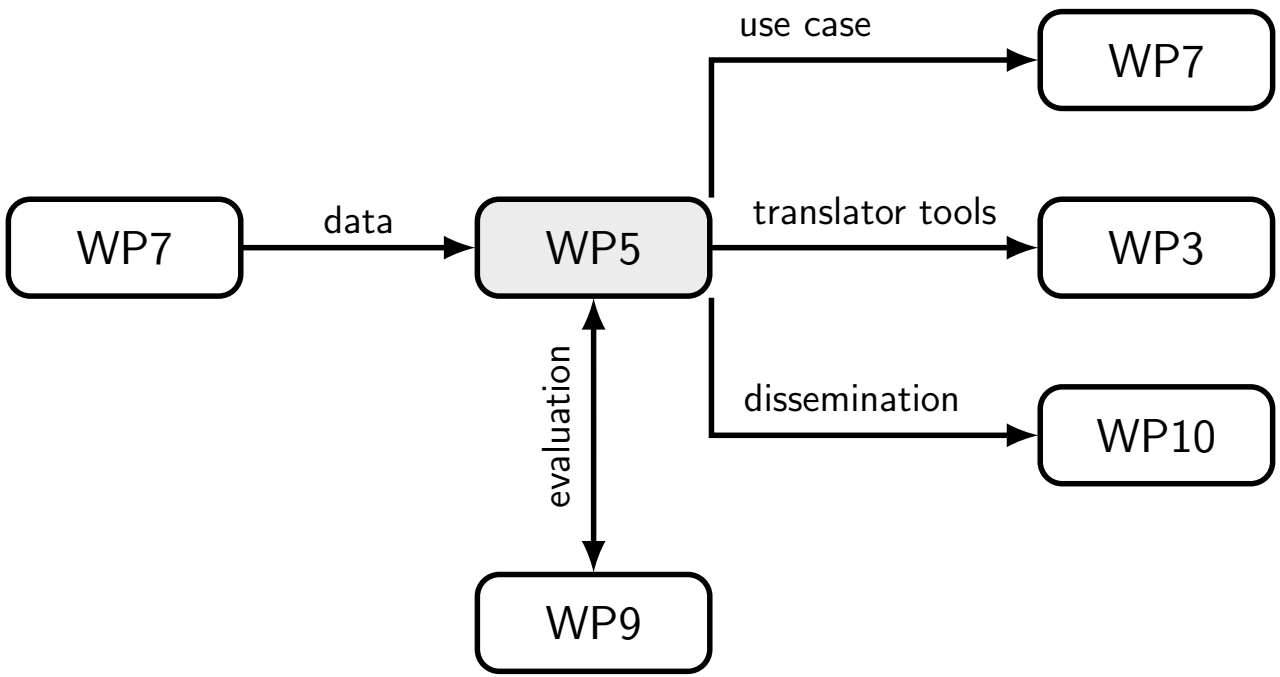
- Should be better and better integrated than an external software like Genia

Krasimir!

Future work

Time to share with other WPs

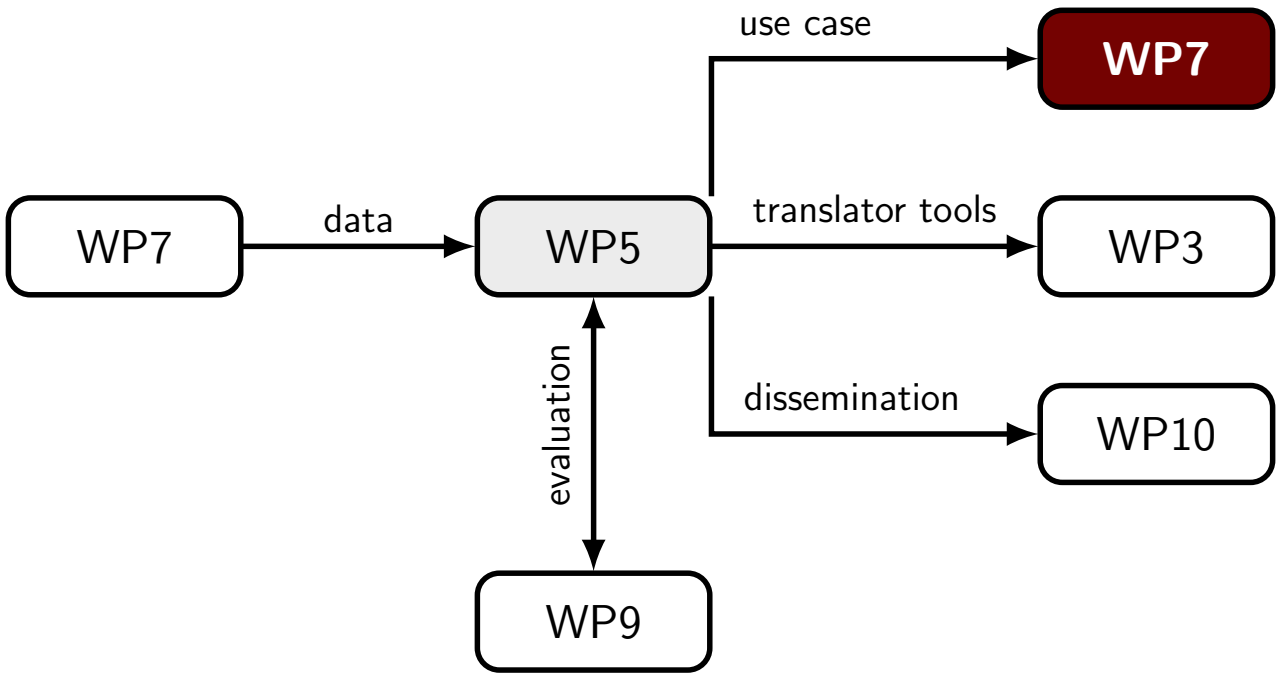
WP5 relations



Future work

Time to share with other WPs

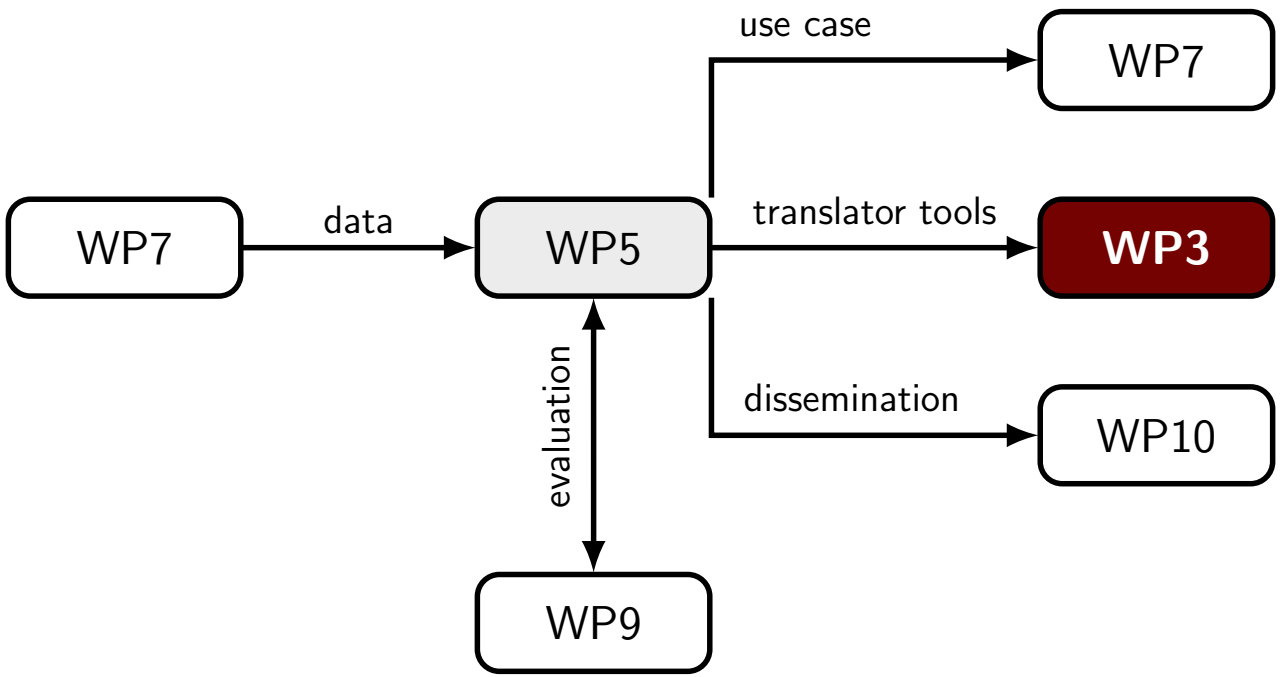
Best system as final translator in the prototype



Future work

Time to share with other WPs

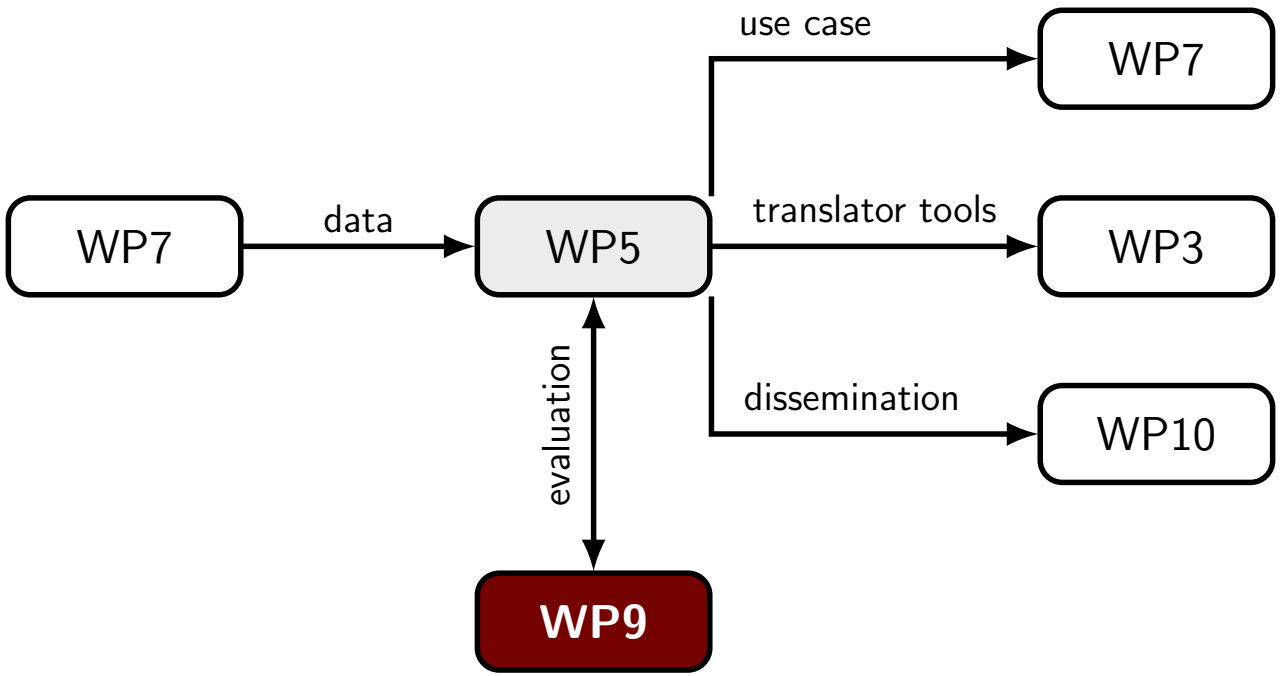
To be included in the translator tools?



Future work

Time to share with other WPs

Manual evaluation of a selection of systems?



Dissemination

Overview

Dissemination

WP dissemination

Refereed Conferences

- *How Much do Grammars Leak? A Grammar Evaluation*
  Krasimir Angelov
  Submitted to COLING (2012).
- *Probabilistic Robust Parsing with Parallel Multiple Context-Free Grammars*
  Krasimir Angelov
  Submitted to COLING (2012).

Dissemination

Toyota's advice

WP5
Statistical and Robust Translation

Cristina España-Bonet

Universitat Politècnica de Catalunya, TALP Research Center

– 5th project meeting –

Utrecht, September 20th, 2012

Ongoing work: Improving the grammar (French)

Example

Version 1
Une composition pharmaceutique `comprising an aqueous solution` **d'arginine et d'ibuprofène**, dans laquelle le rapport molaire **d'arginine à l'ibuprofène** `is less then 1:1`.

Version 2
Une composition pharmaceutique **comprendant une solution aqueuse** `of arginine and ibuprofen`, dans laquelle le rapport molaire `of arginine to ibuprofen is less then 1:1`.

Version 3
Une composition pharmaceutique comprendant une solution aqueuse **d'arginine et de ibuprofen**, dans laquelle le rapport molaire **d'arginine à le ibuprofen** `is less then 1:1`.

Extra slides

Ongoing work: Improving the grammar (French)

Chunk division

	Version 1	Version 2	Version 3
NP	2,366 (52.2%)	2,282 (55.7%)	2,954 (72.1%)
VP	275 (5.7%)	366 (4.4%)	596 (37.2%)
AdjP	82 (36.8%)	63 (42.6%)	107 (100%)
AdvP	1,960 (50.3%)	1,825 (47.0%)	2,439 (62.9%)
RelP	648 (88.3%)	594 (81.4%)	616 (84.5%)
Sum	5,301 (48.2%)	5,130 (49.1%)	6,712 (64.2%)
Total	11,002	10,456	10,456