Published on Multilingual Online Translation (http://www.molto-project.eu)

D9.2 MOLTO evaluation and assessment report

Contract No.:	FP7-ICT-247914
Project full title:	MOLTO — Multilingual Online Translation
Deliverable:	D9.2 MOLTO evaluation and assessment report
Security (distribution level):	Public
Contractual date of delivery:	M38
Actual date of delivery:	2013-05-31 (v1.0)
Type:	Report
Status version:	v1.0
Author(s):	Jussi Rautio, Maarit Koponen
Task responsible:	UHEL
Other contributors:	UZH, UPC

Abstract

This final report describes the evaluation of translation quality in five MOLTO use-cases that implement the Grammatical Framework (GF) for multilingual text generation and translation. Evaluations were made by using both automatically calculated metrics and manually by volunteers.

Contents

List of Figures

List of Tables

1 Introduction

The purpose of this work is to report the automatic and manual quality evaluation results of the translations produced by various MOLTO work packages.

The evaluation of the multilingual translation quality of the GF translations has several goals:

1. to evaluate the accuracy, grammaticality and information transfer of the sentences in multiple languages;
2. to collect information about the amount of effort to manually edit the machine translation suggestions into an acceptable translation;
3. to compare these results with publicly available machine translation systems, like Google Translate, Microsoft Bing and Systran;
4. to find out any remaining issues with the GF grammars in order to fix them (diagnostic evaluation).

To achieve these goals, both automatic metrics and human evaluations were used. The evaluation materials and methods are described in more detail separately for each use case. The source language was English in all of the cases.

2 Methods and tools used in the evaluation

2.1 Appraise

The manual evaluation was carried out mainly on Appraise [5], a web-based open-source platform for the evaluation of machine translation (MT). Appraise allows various MT evaluation tasks like the comparison of two MT systems with each other, post-editing of translation suggestions, and the ranking of suggestions on a chosen scale. Appraise also measures the time each evaluator uses with each example.

2.2 Evaluators

All in all, 45 people participated in the human evaluation of the materials. All evaluators were volunteers, except for the patents reviewers, who were professional translators or proof-readers familiar with patent texts. All evaluators were native or near-native speakers of the respective target language. If possible, the same persons evaluated the tourist phrasebook, ACE-in-GF and Cultural heritage material to keep the results consistent. The mathematics use cases were evaluated by people familiar with the language and conventions of mathematic writing.

The evaluators were not told which MT systems were used to create the translation suggestions, and the order of the MT suggestions by each MT system was randomized by the evaluation tool for each evaluated sentence. If the TM suggestion required post-editing, the evaluators were instructed to make as few edits as possible. If none of the translation suggestions were acceptable, the evaluators were also able to translate the sentence completely from scratch.

2.3 Metrics used in the evaluation

All the automatic metrics were calculated using the open-source Asiya toolkit [6] by comparing the output of the MT systems to the reference translations. Both lexical and edit distance based metrics were used. In the Patents use case, human translations produced from scratch were used as reference, while post-edited machine translations were used as reference in the tourist phrasebook and Ace-in-GF cases.

2.3.1 Lexical metrics

BLEU (Bilingual Evaluation Understudy) [9]: BLEU is based on n-gram matching and measures the lexical precision of the MT output with the reference translations. The BLEU metric works well on a corpus level, but is less reliable on a sentence level. BLEU is the standard metric in the comparison, evaluation and development of MT systems.

NIST [3]: NIST score is based on BLEU but with modified brevity penalty and information weighting. In calculating the information weight, less frequently occurring n-grams are considered more informative and thus given more weight.

2.3.2 Edit distance based metrics

WER (Word error rate) [8]: WER calculates the edit distance between the MT and reference translation as the number of word-level insertions, deletions and substitutions divided by the number of words in the reference. Changes in word order are treated as insertions and deletions.

PER (Position-independent word error rate) [7]: PER calculates edit distance between between the MT and reference translation similar to WER but allows the matched words to appear in different order. Word order changes are not counted as edit operations, and the score is calculated as the number of word-level insertions, deletions and substitutions divided by the number of words in the reference

TER (Translation Edit Rate) [10]: TER differs from WER and PER by adding a new edit operation type shift, which calculates the number of word order changes. The TER score is calculated as the number of word-level insertions, deletions, substitutions and shifts divided by the number of words in the reference. TER is the generally used metric in post-editing studies (often called HTER when post-edited MT is used as reference). TER quite accurately measures the effort needed to post-edit a translation suggestion into a correct translation.

3 Evaluation of the tourist phrasebook

The first use-case evaluated was the tourist phrasebook material. This phase was also used to test the stability of the Appraise platform and familiarize the evaluators with its use. A pilot test for Finnish with eleven students of translation studies as the evaluators was also carried out in this phase.

Thirteen European languages – Bulgarian, Catalan, Danish, Dutch, Finnish, French, German, Italian, Norwegian, Polish, Romanian, Spanish and Swedish – were evaluated with a sample of 139 test sentences containing 827 English source words. The tourist phrasebook consists of simple phrases and questions with a simple lexicon. The material contained sentences for asking for directions, buying things, questions about relatives and other small-talk (see example below).

3.1 Evaluation sample for the tourist phrasebook

The sample for this evaluation was created by first randomly generating a large set of parse trees using the GF phrasebook grammar. Then the sentences with repeated terminals like 'Do you want to eat a pizza and a warm pizza?' were automatically removed from this set. The remaining parse trees were linearized into the translations of the target languages, and a random sample was selected. The English sentences were translated into the 13 target languages using Google Translate[1] and Bing Translator[2] via their respective Web pages. Swedish, Spanish, Italian, French, German, Dutch and Polish were also translated with SYSTRAN[3], as these are the languages in the test set that SYSTRAN supports.

The disambiguated GF linearization for English was used as the example sentence to avoid any ambiguity with different forms of pronouns, like between familiar and polite forms of 'you':

Are your (singular,polite,female) children married?

[1] `http://translate.google.com`
[2] `http://www.bing.com/translator`
[3] `http://www.systranet.com/translate`

Can you (singular,polite,male) wait for us (male) at the nearest cinema?

They (male) don't speak Flemish.

3.2 Evaluation process for the tourist phrasebook

At least two native speakers of each language evaluated the translation suggestions in Appraise. The examples were presented randomly, and the order of the three or four (in case of SYSTRAN-supported languages) translation suggestions were also randomized, so the evaluators could not know the system that had created each suggestion. The post-editing interface of Appraise is shown in Figure 1 on page 3. The evaluators were told to pay special attention to syntactical and morphological correctness.

The task of the evaluators was to select the translation they to be the best quality and then accept it as-is or post-edit into a good quality translation with minimal amount of editing. If all the suggestions were unacceptable, the evaluator also could create a new translation from scratch. The Appraise system recorded each chosen suggestion and the amount of time taken with the choosing and post-editing. The task took approximately 20–45 minutes per evaluator.

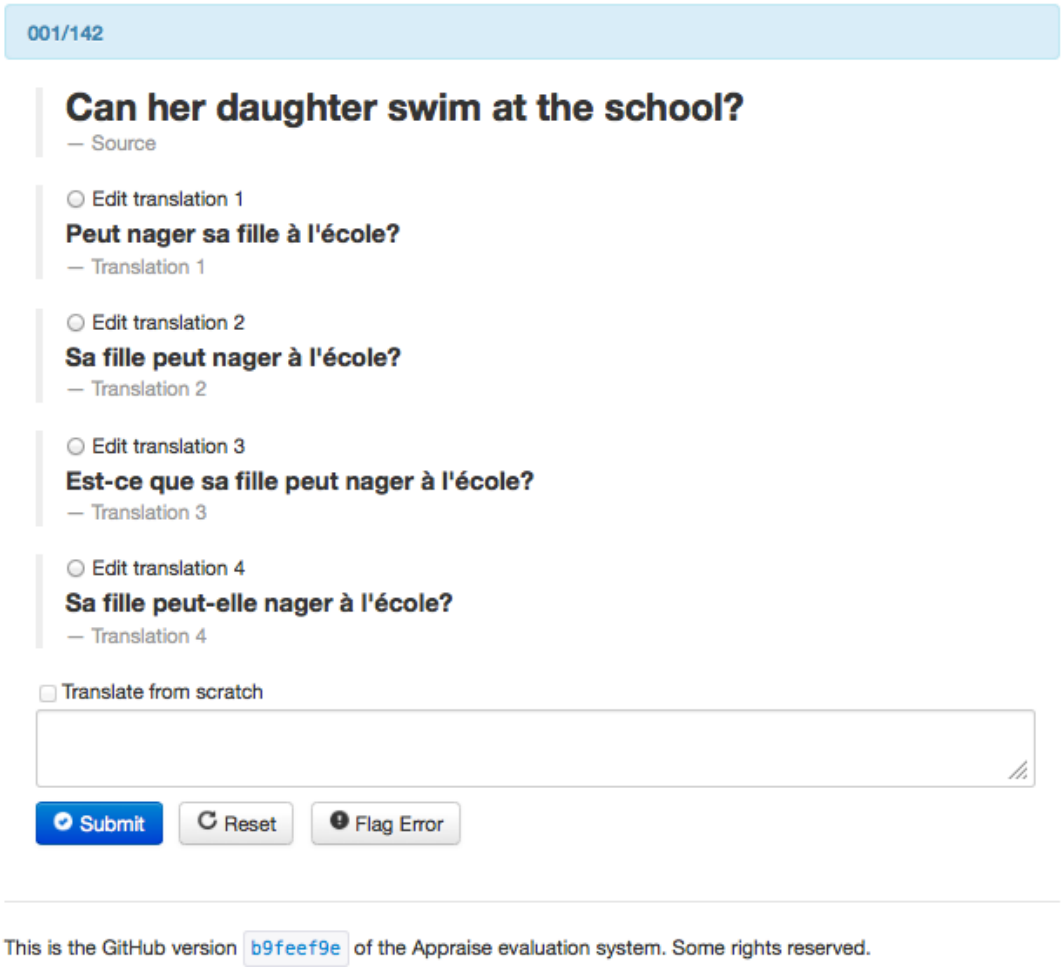

Figure 1:
Appraise post-editing page

3.3 Evaluation results for the tourist phrasebook

3.3.1 Automatic metrics for the tourist phrasebook

As the evaluators accepted as or modified the suggestions into translations they preferred, they created reference translations. These references were then used to calculate the automatic quality metrics for each translation system. The results are presented in Table 1. The best score of each system in each metric is shown in bold – in BLEU and NIST scores higher values are better, and lower values in the edit distance based metrics. BLEU, TER, WER and PER scores always vary between 0 and 1.

Table 1: Automatic metrics for the Tourist phrasebook

	Grammatical Framework					**Google Translate**				
	BLEU	**NIST**	**TER**	**WER**	**PER**	**BLEU**	**NIST**	**TER**	**WER**	**PER**
Bulgarian	0.585	7.172	0.232	0.374	0.252	0.443	6.136	0.331	0.414	0.313
Catalan	0.904	9.027	0.040	0.197	0.148	0.554	6.477	0.309	0.434	0.378
Danish	0.760	8.293	0.100	0.145	0.132	**0.684**	7.703	**0.163**	**0.308**	**0.265**
Dutch	0.814	8.610	0.083	0.125	0.106	0.669	**7.801**	0.176	0.341	0.270
Finnish	0.887	8.373	0.053	0.119	0.099	0.436	5.659	0.330	0.416	0.357
French	0.875	**9.508**	0.078	0.368	0.308	0.625	7.264	0.251	0.449	0.375
German	0.862	8.931	0.052	0.149	0.112	0.523	6.829	0.262	0.406	0.323
Italian	0.902	8.855	0.050	0.223	0.207	0.562	6.630	0.290	0.452	0.379
Norwegian	0.810	8.420	0.085	0.273	0.248	0.575	7.013	0.220	0.369	0.312
Polish	0.801	8.448	0.094	0.168	0.094	0.492	6.370	0.303	0.381	0.307
Romanian	0.690	7.714	0.175	0.307	0.273	0.543	6.563	0.287	0.472	0.407
Spanish	0.926	9.309	0.035	0.135	0.093	0.570	6.694	0.242	0.361	0.319
Swedish	**0.961**	9.349	**0.017**	**0.048**	**0.044**	0.613	7.130	0.194	0.356	0.326
Average	**0.829**	**8.616**	**0.084**	**0.202**	**0.163**	0.561	6.790	0.258	0.397	0.333
	Bing Translator					**SYSTRAN**				
	BLEU	**NIST**	**TER**	**WER**	**PER**	**BLEU**	**NIST**	**TER**	**WER**	**PER**
Bulgarian	0.398	5.715	0.331	0.443	0.330					
Catalan	0.535	6.567	0.292	0.444	0.381					
Danish	0.615	7.499	**0.174**	0.328	0.295					
Dutch	0.580	7.333	0.226	0.376	0.282	0.647	**7.445**	**0.174**	0.348	0.294
Finnish	0.369	5.239	0.377	0.451	0.397					
French	0.637	7.444	0.252	0.433	0.371	**0.649**	7.344	0.201	0.433	0.380
German	0.607	7.304	0.209	0.361	0.288	0.606	7.045	0.227	0.378	0.325
Italian	**0.688**	7.392	0.203	0.355	0.307	0.568	7.004	0.237	0.399	**0.286**
Norwegian	0.529	6.788	0.248	0.390	0.322					
Polish	0.423	5.927	0.340	0.412	0.357	0.196	4.030	0.526	0.598	0.529
Romanian	0.434	5.877	0.349	0.483	0.412					
Spanish	0.675	**7.536**	0.176	**0.311**	**0.249**	0.593	7.081	0.202	**0.344**	0.298
Swedish	0.532	6.700	0.235	0.375	0.340	0.485	6.136	0.290	0.436	0.384
Average	0.540	6.717	0.263	0.397	0.333	0.535	6.583	0.265	0.419	0.357

As the average scores show, the GF translations got the best score in each metric, with Google Translate, Bing Translator receiving nearly identical scores in all of the metrics.

3.3.2 Evaluator preferences for the tourist phrasebook

The MT system that the evaluators preferred for the most accurate translation was also recorded. The evaluators could either:

1. Accept the GF translation or the exact same translation from another system without any editing (a perfect match)

2. Accept the GF translation or the exact same translation from another system for post-editing

3. Reject the GF translation by accepting a translation suggestion of another MT system as such or for post-editing

4. Reject all the given MT suggestions and translate the example sentence from scratch.

Translating from scratch was used very rarely, as it is usually easier to re-use parts of even a poor-quality suggestion than to rewrite the whole sentence. There was only one instance in the whole evaluation sample in which both evaluators chose to translate the example from scratch.

The evaluator preferences are shown in Table 2 and in Figure 2. Between all the languages 28 percent for Bulgarian to 91 percent for Swedish of the GF translations were accepted as such by both or one of the evaluators, and only 6–28 percent of the translation suggestions were selected from any of the other systems or translated from scratch. On average, 84 percent of the GF translations were preferred as such of a candidate for post-editing.

Table 2: Evaluator preferences for the Tourist phrasebook

	Accepted GF				**Post-edited GF**				**Other MT/from scratch**	
	Both	**One**	**Total**	**Acc. %**	**Both**	**One**	**Total**	**PE %**	**Both**	**Total**
Bulgarian	13	26	39	**28 %**	24	37	61	**44 %**	39	**28 %**
Catalan	39	72	111	**80 %**	9	8	17	**12 %**	11	**8 %**
Danish	49	30	79	**57 %**	14	17	31	**22 %**	29	**21 %**
Dutch	22	74	96	**69 %**	6	0	6	**4 %**	37	**27 %**
Finnish	73	34	107	**77 %**	17	5	22	**16 %**	10	**7 %**
French	55	49	104	**75 %**	2	18	20	**14 %**	15	**11 %**
German	55	47	102	**73 %**	16	0	16	**12 %**	21	**15 %**
Italian	73	40	113	**81 %**	0	4	4	**3 %**	22	**16 %**
Norwegian	41	44	85	**61 %**	20	10	30	**22 %**	24	**17 %**
Polish	66	25	91	**65 %**	23	9	32	**23 %**	16	**12 %**
Romanian	25	41	66	**47 %**	15	24	39	**28 %**	34	**24 %**
Spanish	69	46	115	**83 %**	4	1	5	**4 %**	19	**14 %**
Swedish	80	47	127	**91 %**	1	3	4	**3 %**	8	**6 %**
Average	**50.8**	**44.2**	**95.0**	**68 %**	**11.6**	**10.5**	**22.1**	**16 %**	**21.9**	**16 %**

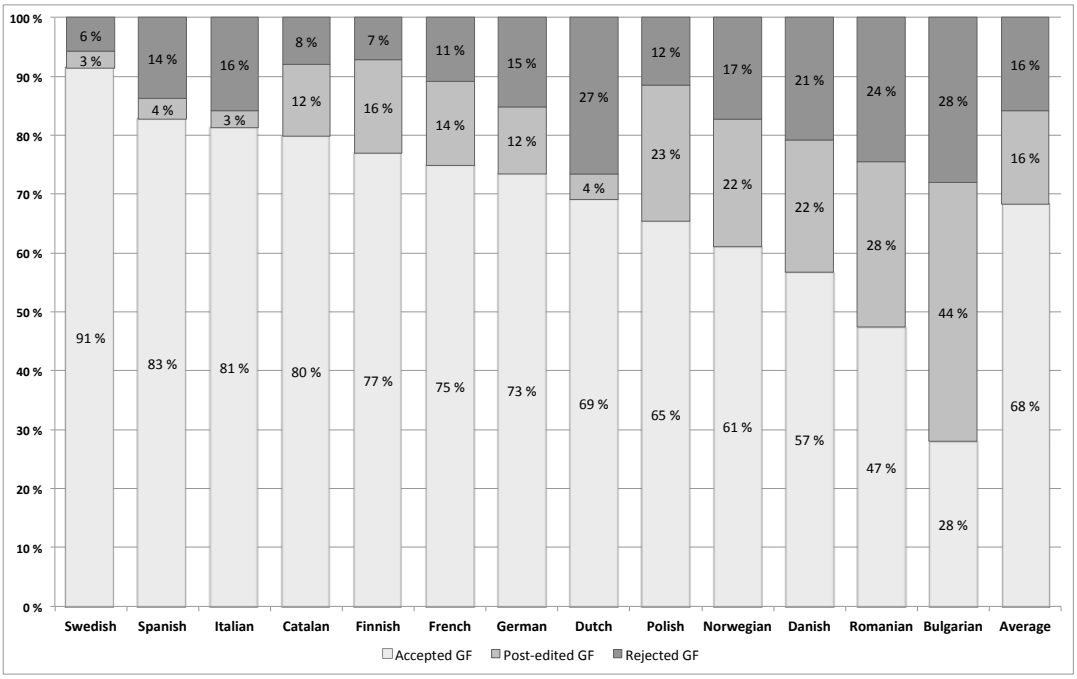

Figure 2: Evaluator preferences in the tourist phrasebook

3.4 Discussion of the phrasebook results

Some recurring issues in human evaluation were observed. As the phrasebook sample was the first one to be evaluated, the evaluators were not yet familiar with the evaluation process or the tool. For example, one Italian evaluator chose to translate most of the samples from scratch, even if nearly identical translation suggestions were available. This can be seen in the Italian results as a higher number of rejected GF translations.

Most of the corrections the evaluators made were lexical, like ‘abitano’ to ‘vivono’ in Italian ‘to live’. Some corrections were – like in all human evaluations – matters of preference. For example, the Finnish evaluators used the semantically similar translations ‘auki/avoinna’ for ‘open’ and ‘kiinni/suljettu’ for ‘closed’, whereas GF uses the forms ‘avoinna’ and ‘kiinni’ consistently.

Another strength of the GF translations was in the use of polite and informal forms of ‘you’. As the example translations with ‘you’ in them were translated with Google Translate, Bing Translator and SYSTRAN, the systems returned either the polite or the informal form quite randomly. With GF, it is easy to disambiguate the correct forms in the original parse tree.

The relatively low automatic metric and acceptance scores for Bulgarian and Romanian were found to be caused by easily fixable features in the resource grammars: Both languages are pro-drop languages, but GF uses the construction with a pronoun in its translations. Also, Bulgarian has both clitic and non-clitic forms of personal pronouns as in Romance languages, and GF implements only the non-clitic forms. The GF translations of Bulgarian and Romanian were therefore technically grammatical, but those were not preferred by the evaluators.

As both the automatic metrics and the evaluators’ preferences show, GF translations scored significantly better than the other MT systems evaluated. The is also a good correlation between the automatic metrics and the human evaluation results as seen in Figure 3 on page 7. The very low average TER score of 0.084 for the GF translations shows that the GF suggestions required much less effort to post-edit into correct translations.

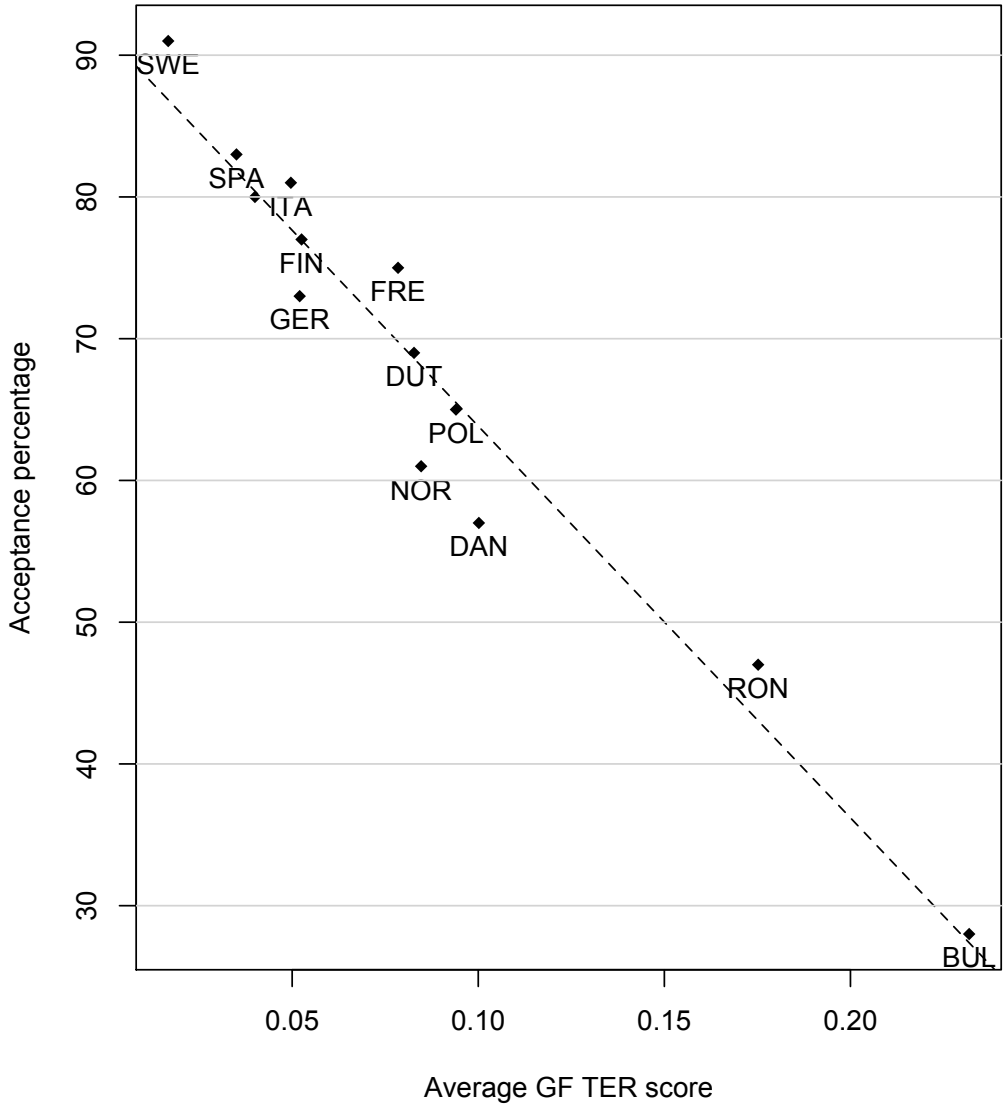

Figure 3: Correlation between the average TER score and acceptance percentage ($c = -0.973$)

4 Evaluation of ACE-in-GF

4.1 Evaluation sample for ACE-in-GF

The ACE-in-GF evaluation material comprises a total of 111 sentences having 3–18 words each (848 words of English source text in all). A detailed description of the methods used for generating and selecting the example sentences is given in Deliverable 11.3 [1]. The main goal of the sample selection was to create a set that covered the non-lexical functions of the ACE grammar as completely as possible. Examples of the sentences evaluated are:

Every dog is a cat or is a horse.

Nothing that is a dog or that is a cat is a bird and is a horse.

What doesn't John hate?

Which cat hates John and pushes a tail?

Th ACE-in-GF evaluation was carried out from English into Catalan, Danish, Dutch, Finnish, French, German, Italian, Norwegian, Spanish and Swedish. Other languages included in the grammar, namely Bulgarian, Greek, Latvian, Polish, Russian, and Romanian, were not evaluated, either because suitable evaluators were not available or because the grammar still lacked the implementation of some important constructs, e.g. 'if-then', at the time of the evaluation. English was used as the source language to be evaluated against. For comparison purposes, the English sentences were also machine translated using Google Translate.

Like in the phrasebook evaluation, two native speakers of each target language were recruited for the evaluation – the same people that evaluated the phrasebook material were preferred. The evaluators were not familiar with CNL or ACE and were not told that the translations were automatically created or introduced to the involved translation technologies beforehand. The evaluators were presented with a source sentence in English and in this round, only two translation options, ACE-in-GF or Google Translate. Again, the task was to choose the translation result they considered best and either accept it as-is, post-edit it as necessary or translate the example from scratch.

4.2 Evaluation results for ACE-in-GF

4.2.1 Automatic metrics for ACE-in-GF

The system-level automatic metrics calculated from the sample are presented in Table 3. All metrics used measure the lexical level similarity of the translation suggestions and the reference translations.

Table 3: Automatic metrics: ACE-in-GF vs. Google Translate

	ACE-in-GF					**Google Translate**				
	BLEU	**NIST**	**TER**	**WER**	**PER**	**BLEU**	**NIST**	**TER**	**WER**	**PER**
Catalan	0.809	8.803	0.101	0.231	0.223	0.716	7.993	0.151	0.265	0.232
Danish	0.716	8.233	0.142	0.263	0.208	0.623	7.452	0.186	0.324	0.244
Dutch	0.899	9.335	0.056	0.223	0.158	0.735	8.371	0.133	0.275	**0.170**
Finnish	**0.948**	**9.336**	**0.026**	**0.147**	**0.132**	0.446	6.053	0.321	0.401	0.365
French	0.873	8.998	0.073	0.221	0.179	0.784	8.284	0.128	0.258	0.217
German	0.850	9.027	0.060	0.162	0.152	0.660	7.943	0.166	0.289	0.187
Italian	0.822	8.626	0.090	0.191	0.173	0.793	8.186	0.116	**0.204**	0.181
Norwegian	0.718	8.142	0.116	0.248	0.187	0.687	7.795	0.152	0.240	0.199
Spanish	0.788	8.835	0.095	0.224	0.198	0.708	7.994	0.167	0.281	0.212
Swedish	0.889	9.303	0.056	0.300	0.226	**0.794**	**8.723**	**0.093**	0.260	0.194
Average	**0.831**	**8.864**	**0.081**	**0.221**	**0.184**	**0.695**	**7.879**	**0.161**	**0.280**	**0.220**

Like with the phrasebook sample, all average scores for ACE-in-GF translations are better than the respective results for Google Translate. ACE-in-GF gets the best scores with Finnish, while Google Translate fares the worst. As far as ACE-in-GF is concerned, this is not surprising as the Finnish concrete syntax in ACE-in-GF — together with German and Spanish — received more developer attention than the other languages.

4.2.2 Human evaluation for ACE-in-GF

The translation suggestion that the evaluators preferred is presented in Table 4 on page 9 and Figure 4 on page 9. For example, 87% of the ACE-in-GF translation suggestions in Finnish were accepted without editing by both or one of the evaluators, and 10% was chosen for post-editing. Only 3% of the Finnish Google suggestions were preferred as such or for post-editing and the ACE-in-GF suggestion rejected.

Table 4: Evaluator preferences in ACE-in-GF translations

	Accepted ACE-in-GF				Post-edited ACE-in-GF				Pref. Google	
	Both	**One**	**Total**	**Acc. %**	**Both**	**One**	**Total**	**PE %**	**Either**	**Total**
Catalan	40	22	62	**56 %**	9	20	29	**26 %**	20	**18 %**
Danish	29	19	48	**43 %**	23	12	35	**32 %**	28	**25 %**
Dutch	44	38	82	**74 %**	9	3	12	**11 %**	17	**15 %**
Finnish	71	26	97	**87 %**	11	0	11	**10 %**	3	**3 %**
French	47	30	77	**69 %**	7	2	9	**8 %**	25	**23 %**
German	57	18	75	**68 %**	13	5	18	**16 %**	18	**16 %**
Italian	45	16	61	**55 %**	0	8	8	**7 %**	42	**38 %**
Norwegian	32	20	52	**47 %**	20	11	31	**28 %**	28	**25 %**
Spanish	27	25	52	**47 %**	27	26	53	**48 %**	6	**5 %**
Swedish	32	48	80	**72 %**	4	13	17	**15 %**	14	**13 %**
Average	**42.4**	**26.2**	**68.6**	**61.8 %**	**12.3**	**10.0**	**22.3**	**20.1**	**20.1**	**18.1 %**

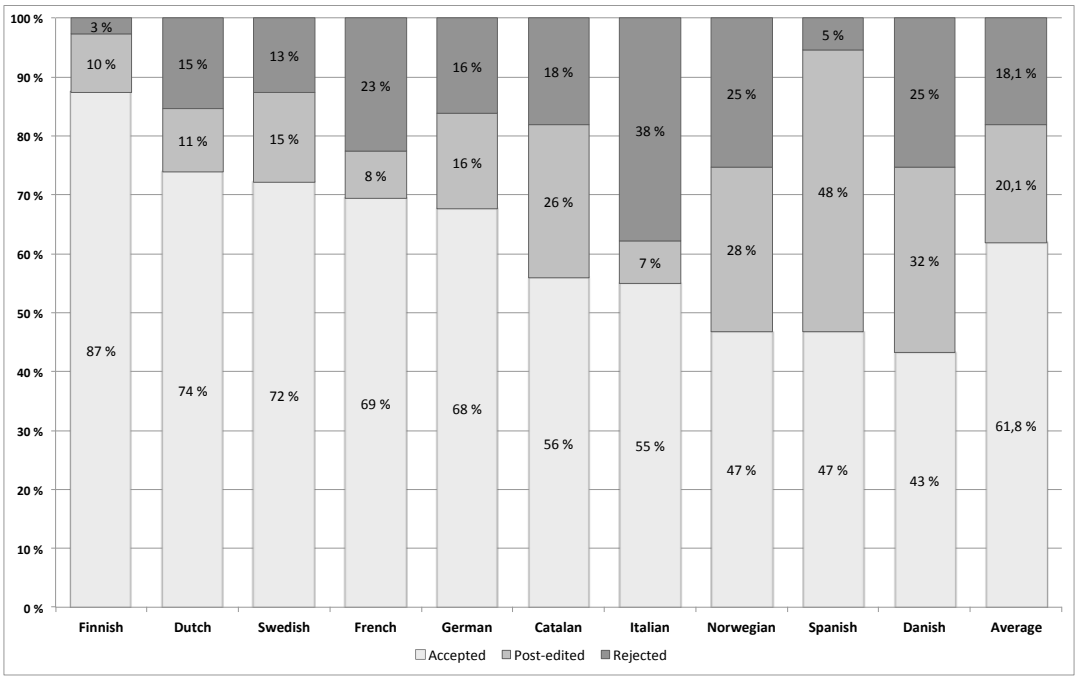

Figure 4: Evaluator preference in ACE-in-GF

4.3 Issues in human evaluation of ACE-in-GF

The evaluators had some difficulties with certain issues in the ACE-in-GF translations. Even though the evaluators were instructed to ignore the content words of the sentences and only focus on the syntax and morphology, some evaluators found the material hard to evaluate. For example,the lack of an ellipsis in the source sentences generated by ACE-in-GF was seen as a flaw in the translations. For example the sentence 'Everything that something finds is a horse or is a bird.' was usually translated with a more natural-sounding ellipsis 'Everything that something finds is a horse or (is) a bird.' This had a negative effect on the scores of the GF translations, as Google Translate suggestions usually used an ellipsis in its translations.

As with all manual evaluation of translations, some choices made by the evaluators were purely subjective, for example the use of punctuation and using the active form instead of the passive. For example, the question 'What is seen by every dog?' was translated into 'What does every dog see?' by the evaluators in many cases.

Even with these issues, it is evident that ACE-in-GF translation suggestions were again preferred over

the Google Translate ones. In line with the phrasebook score, a very low TER score of 0.081 (0.084 in the phrasebook) again shows that very little post-editing is needed to create high-quality translations.

4.4 Combined results for the phrasebook and ACE-in-GF evaluations

Figure 5 on page 10 shows the combined preferences for the phrasebook and ACE-in-GF material for the ten languages evaluated in both cases.

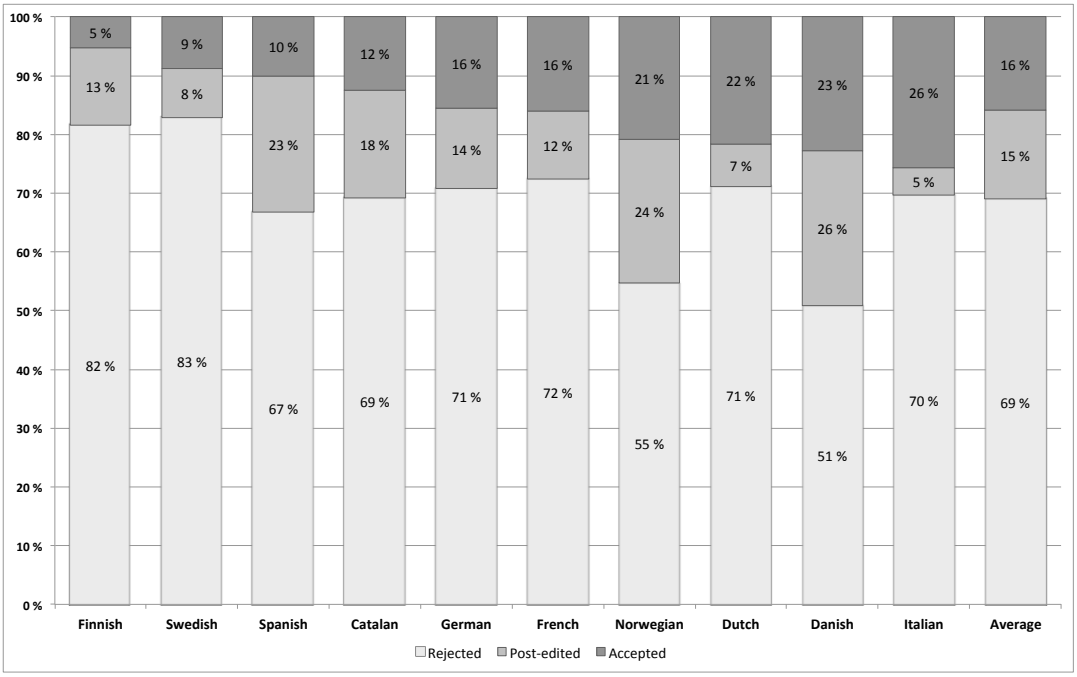

Figure 5: Combined evaluator preferences in phrasebook and ACE-in-GF results

5 Evaluation of the patents

The patents use-case in the MOLTO project implements GF technologies in translation of open-domain texts. The patent sample contains the abstracts and claims parts of three different biomedical patents referred in this and other deliverables as EPOA61P, PATSA61P and USAPATS.

The PATSA61P sample is from the MAREC (MAtrixware REsearch Collection)[4] corpus, which is a data collection over 19 million of European patents made available by the CLEF Initiative.

The EPOA61P sample is from EPO (European Patents Office)[5] corpus.

The USAPATS sample contains patents from the United States Patent and Trademark Office (UPSTO)[6]. This set was selected because the EPO patent corpus might be used as Google's training data and therefore give Google Translate better results with the PATSA61P test case. This set has no human-made reference translations, so the automatic metrics could not be calculated.

5.1 Evaluation sample for the patents

The evaluation set for the patents was selected from the three patent cases by first translating the source sentences with a phrase-based SMT system trained on the biomedical domain and a hybrid system (in this report referred to as 'GF hybrid') based on this SMT system, that also makes use of a static lexicon for GF, and does a soft integration between SMT phrases and GF multiple translation options. A detailed report of the MT systems and the training corpora are presented in Deliverable 5.3 [4]. The source sentences were also translated with Google Translate and SYSTRAN systems. In the USAPATS sample the PLuTO (Patent Language Translations On-line)[7] translations were also used in the comparison.

The translations were then examined, and all source sentences with two or more exactly same translations were removed; this was done to not make the evaluators use time to look for differences in the translation suggestions when there would be none. Extremely long examples were also removed, as there were examples of sentences containing over 2000 words, usually names of chemical compounds. Finally, a random sample of each patent case was selected. The size of the resulting sample sets are shown in Table 5.

Table 5: Sample sizes/Total size of the patent texts

	Source	EPOA61P	PATSA61P	USAPATS	Total
German	**Sentences**	200/847	149/1008	75/999	**425/2854 (14.8%)**
	Words	5707/30999	4244/31239	2431/29163	**12382/91491 (13.5%)**
French	**Sentences**	191/858	148/1008	75/999	**423/2865 (14.8%)**
	Words	5788/31964	4224/31239	2431/29163	**12443/92366 (13.5%)**

5.2 Evaluation process for the patents

The evaluators with experience in patents translations was asked to rank each translation suggestion according to the following scale measuring the effectiveness of information transfer[8]:

- **4 = Complete:** All of the information in the source was available from the target; reading the source did not add to information or understanding.
- **3 = Useful:** The information in the target was correct and clear, but reading the source added some additional information or understanding.
- **2 = Marginal:** The information in the target was correct, but reading the source provided significant additions of clarifications.

[4] http://www.ir-facility.org/prototypes/marec
[5] http://www.epo.org
[6] http://www.uspto.gov
[7] http://www.pluto-patenttranslation.eu/
[8] http://www.taus.org

- **1 = Poor:** The information in the target was unclear and/or incorrect; reading the source would be necessary for understanding.

Also, a translation suggestion with missing or untranslated words was never given the rank of 4. The evaluators were also asked to pay attention to the conventions and standards of patent translations.

All three cases were completely ranked by one German and one French evaluator. Another German evaluator ranked the whole USAPATS sample and half of the EPOA61P and PATSA61P samples.

The evaluation was done in the ranking interface of Appraise (See Figure 6 on page 12). Again, the suggestions by different MT systems were presented in random order. It took each evaluator about 6–7 hours to evaluate all the sentences.

Figure 6: The ranking interface of Appraise

5.3 Evaluation results for the patents

5.3.1 Automatic metrics for the patents

The automatic BLEU and TER metrics for EPOA61P and PATSA61P samples were calculated using the Asiya toolkit with the human reference translations provided. The results are presented in Table 6 on page 13 and in Figure 8 on page 14 and in Figures 7 and 8 on page 14.

Table 6: Automatic metrics for the EPOA61P and PATSA61P samples

	GF hybrid		**SMT**		**Google**		**SYSTRAN**	
German	**BLEU**	**TER**	**BLEU**	**TER**	**BLEU**	**TER**	**BLEU**	**TER**
EPOA61P	0.465	0.356	0.474	0.341	0.502	0.334	0.185	0.649
PATSA61P	0.468	0.351	0.476	0.346	0.507	0.321	0.195	0.642
French								
EPOA61P	0.592	0.268	0.600	0.260	0.602	0.255	0.332	0.458
PATSA61P	0.563	0.294	0.579	0.264	0.583	0.253	0.338	0.432

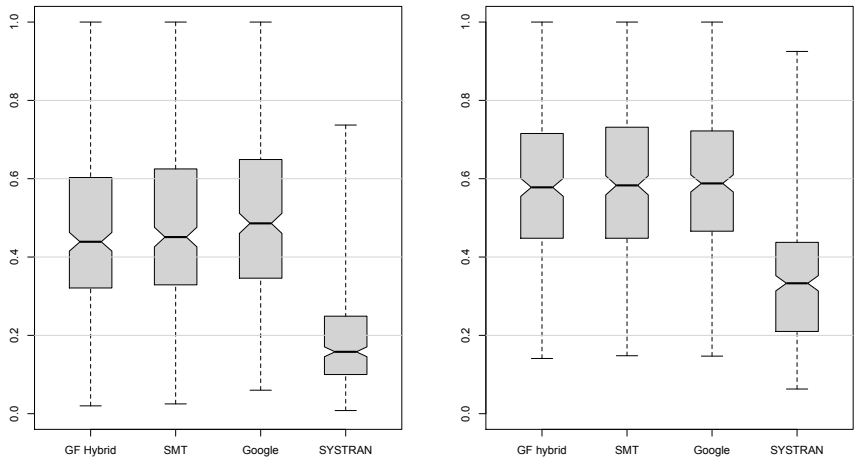

Figure 7: Boxplots for combined BLEU scores of EPOA61P and PATSA61P for German (left) and French (right)

As the metrics show, there is no significant difference between the results for EPOA61P and PATSA61P either in German and French or between the GF hybrid, the SMT system and Google Translate.

5.3.2 Human rankings for the patents

The Table 7 on page 14 and Table 8 on page 14 show the averages and medians of the rankings of each evaluation sample. Appendix A on page 21 shows the rankings of each patent case individually.

The human rankings show no significant difference between the translations of the GF hybrid system and the SMT system in either language. Surprisingly, the rankings for the PLuTO system vary greatly in German USAPATS patent sample: Evaluator 1 ranks PLuTO as the best system overall, but Evaluator 2 ranks it as the second worst. This can be clearly seen in Figure 11 on page 22, where Evaluator 2 gave ten times as many lowest rankings of 1 to PLuTO as Evaluator 1.

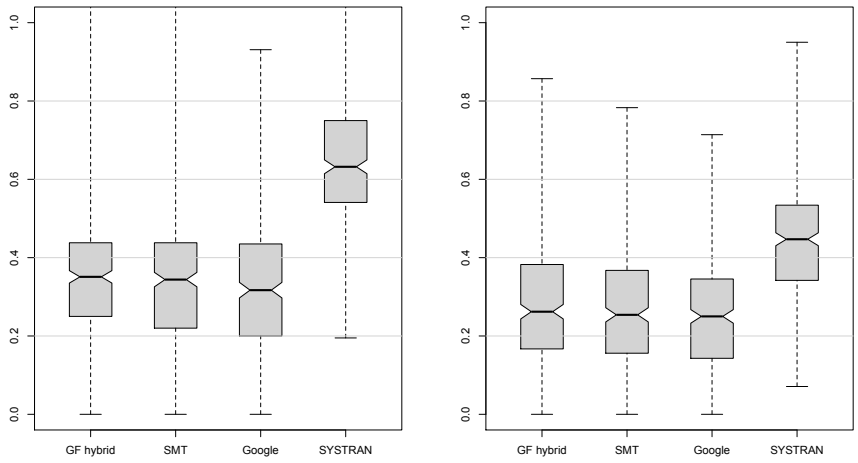

Figure 8: Boxplots for combined TER scores of EPOA61P and PATSA61P for German (left) and French (right)

Table 7: Rankings for the German patent samples

		German Evaluator 1				
		GF hybrid	**SMT**	**Google**	**Systran**	**PLuTO**
PATSA61P	Average	3.07	3.09	**3.22**	2.45	n/a
	Median	**3**	**3**	**3**	2	n/a
EPOA61P	Average	2.65	2.80	**2.90**	2.085	n/a
	Median	**3**	**3**	**3**	2	n/a
USAPATS	Average	1.74	1.67	1.85	1.18	**2.53**
	Median	2	2	2	1	**3**
		German Evaluator 2				
		GF hybrid	**SMT**	**Google**	**Systran**	**PLuTO**
PATSA61P	Average	2.22	2.23	**2.30**	1.51	n/a
	Median	**2**	**2**	**2**	1	n/a
EPOA61P	Average	2.49	2.41	**2.58**	1.20	n/a
	Median	2	2	**3**	1	n/a
USAPATS	Average	1.87	1.84	**2.27**	1.39	1.79
	Median	2	2	**3**	1	2

Table 8: Rankings for the French patent samples

		GF hybrid	**SMT**	**Google**	**Systran**	**PLuTO**
PATSA61P	Average	3.09	3.09	**3.28**	2.55	n/a
	Median	3.5	3	**4**	3	n/a
EPOA61P	Average	2.97	2.99	**3.14**	2.07	n/a
	Median	3	3	**4**	2	n/a
USAPATS	Average	2.55	2.52	**3.04**	1.75	3.01
	Median	3	3	3	1	**4**

5.4 Discussion of the patents results

Both the automatic metrics and the human evaluations show that both the SMT system and the GF hybrid based on it perform equally well. The automatic metrics between the GF-based systems and Google Translate are nearly identical. As it is possible that the test cases used in this evaluation are included in Google's training data, evaluation with another new sample not yet used by Google training data might yield even

better results for the GF systems.

Also, the differences between the evaluation sentences by different MT systems were often very trivial and probably hard to always spot. The next example with the differences to the reference translation emphasized shows this:

Reference:

Pharmazeutische Zusammensetzung, umfassend das Salz nach Anspruch 1 und einen pharmazeutisch annehmbaren Träger.

GF hybrid:

Eine pharmazeutische Zusammensetzung, umfassend das Salz *von* Anspruch 1 und einen pharmazeutisch *verträglichen* Träger. (TER 0.200)

SMT:

Pharmazeutische Zusammensetzung, umfassend das Salz *von* Anspruch 1 und einen pharmazeutisch *verträglichen* Träger. (TER 0.133)

Google:

Pharmazeutische Zusammensetzung, umfassend das Salz nach Anspruch 1 und einen pharmazeutisch *verträglichen* Träger. (TER 0.067)

As can be seen, the only difference between the GF hybrid and the SMT system is the indefinite article 'eine', and the difference between both GF systems and Google is the preposition 'von'.

This is also an example of a case where the automatic and manual evaluation scores do not correlate very well. In the previous example, all of these translation suggestions got the highest rank from the evaluator, but the GF hybrid translation got harshly penalized in the TER score for adding one and changing two words.

An issue with evaluation in general is that sometimes even similarity to human translation does not necessarily indicate high quality as assessed by human evaluators. In the patent case, some cases were observed where the evaluators gave poor scores to translations that had good automatic scores, even some where the MT was identical to the human translation generated without MT. This may suggest two things: It may be that the relatively few differences between the MT and human reference are more critical than their number indicates. When sentences identical to the reference are given low quality scores, this indicates that the human references translation would have been scored low, as well. This may relate to the complexity of the sentences, or the fact that the evaluators were evaluating individual sentences without reference to the wider context.

6 Evaluation of the mathematics material

The evaluation set consisted of 107 mathematical clauses generated manually with the MGL grammar. All these clauses were then linearized into German, French and Swedish. A subset of 57 clauses was also linearized into Finnish. Examples of the clauses are:

The proposition that x is equal to y is equivalent to the proposition that y is equal to x

For all x in A, infinity is greater than x

The complex number with polar coordinates 1 and pi is equal to minus 1

The English source example and the translations for French, German, Swedish and Finnish were evaluated by a multi-lingual person with a decades-long experience in mathematical terminology and writing mathematics teaching materials. The evaluator was asked to correct any ambiguities, unconventional on incorrect use of terms and other linguistic errors like errors in word order or morphology found in the samples. All the languages were evaluated simultaneously.

After the evaluation the issues found were found to be in six different categories (modified from the categories in [11]):

Issues making the clause unreadable or ambiguous:

Rewrite needed or information missing
The evaluated sentence was missing some vital information for its correct interpretation or it had to be rewritten from scratch to be understandable.

Incorrect word order
The word order caused an ambiguity or made the clause difficult to understand without reading the source, but no information was missing.

Issues making the clause ungrammatical, but understandable:

Incorrect, unneeded or missing preposition
A preposition was either incorrect unneeded or missing, but this did not impede the interpretation of the clause.

Incorrect case or gender
Nouns, articles or adjectives in the wrong gender or in an incorrect linguistic case.

Incorrect morphology
Words that were inflected incorrectly or had morphological errors

Incorrect terminology
Mathematical terms that were incorrect or not in common use

Each issue was recorded only once per language, though it usually appeared multiple times in the translations. For example, the Finnish genitive form was reversed in many examples. A clause could have issues from one or more categories. The results are presented in Table 9 on page 16.

Table 9: Error categories in mathematics (total 107 examples, 57 in Finnish)

	German	French	Swedish	Finnish
Rewrite needed or information missing	0	0	1	16
Incorrect word order	9	5	8	12
Total	**9**	**5**	**9**	**28**
Incorrect or missing preposition	12	10	5	n/a
Incorrect case or gender	11	6	3	5
Incorrect morphology	0	0	2	5
Incorrect terminology	6	7	8	2
Total	**29**	**23**	**18**	**12**

Almost all German, French and Swedish clauses were evaluated as understandable and unambiguous – this was also reported by the evaluator in a free-form assessment of the translations. None of the German, French or Swedish clauses needed a complete retranslation. The issues in these languages are relatively easy to correct in the lexicon by changing the incorrect prepositions and terms.

On the other hand, nearly half (28 of the 57) of the Finnish clauses were evaluated either as needing a major rewrite or ambiguous. This might be because the MGL grammar is built by GF functors, which means the syntactic structures for all languages are the same, and only the lexicon for each language needs to be created separately. This method seems to work reasonable well for the other languages, but Finnish needs a set of exception rules to perform correctly. For example, as Finnish has a very few prepositions, the prepositional phrases in the other languages need to be presented as relational clauses in Finnish. The Finnish word order may also cause ambiguities, for example in:

The intersection of A and the cartesian product of A and B is a subset of the empty set

A:n ja A:n ja B:n karteesisen tulon leikkaus on tyhjän joukon osajoukko

(The intersection of the cartesian product of A and A and B is a subset of the empty set)

This can be fixed by reordering the clause into

A:n ja B:n karteesisen tulon ja A:n leikkaus on tyhjän joukon osajoukko

These exception rules were being created at the time this report was written.

7 Evaluation of the Cultural heritage material

The Cultural heritage evaluation set contained 48 sample evaluation segments with 51 unique sentences containing 671 words generated by the description and query application grammars presented in Deliverable 8.3 [2] with 33 queries and 15 results translations, for example:

1: Charles, Prince of Wales was painted by Isaac Oliver in 1615.

2: Les demoiselles d'Avignon was painted on canvas by Pablo Picasso in 1937. It is of size 234 by 244 cm and it is painted in red and white. This oil painting is displayed at the Museum of Modern Art.

3: Lady with an Ermine was painted on wood by Leonardo da Vinci in 1490. It is of size 54 by 39 cm.

4: show everything about all miniatures

5: show everything about all miniatures at the Addison Gallery of American Art

6: who painted A Burial At Ornans

Of the 15 languages with application grammars, seven languages were evaluated: Catalan, Danish, Finnish, French, German, Norwegian and Swedish. One to three native speakers of these languages were shown the source sentences and the GF translation and asked to post-edit the suggestions into correct translations. The evaluators were described the situation where the translations were to be used: In querying a database on different qualities of museum objects like material, size or painter and receiving the answers as natural language. The test set was presented as query-result pairs. This evaluation was mainly planned to be a diagnostic evaluation to judge the correctness and information transfer of the sentences created by the grammars.

As the lexicon for the names of the artists, works of art and museums were collected from different ontologies, these names were not localized into the target languages: For example, 'Hans Holbein the Younger' and 'Lady with an Ermine' were not translated into 'Hans Holbein der Jüngere' and 'Dame mit dem Hermelin' in German, and the evaluators were asked to ignore this. The only exception to this was Finnish, in which a morphological case marker was required in the museum names.

7.1 Evaluation results for the Cultural heritage material

Even though at this point of the project the evaluators were familiar with the evaluation methodology, a surprising amount of edits were made into the GF suggestions. As expected from the previous evaluations, Finnish and Swedish had the least amount of changed suggestions by the evaluators, three and one respectively. On average, 38 suggestions for the 51 different sentences were edited in each language.

The high number of sentences edited seems surprising. This is mainly due to the nature of the material: Although each sentence in the evaluation material is unique, the individual sentences are built of a limited set of lexical items and grammatical structures combined in different ways during the generation process.

As all occurrences of the same structure or the same lexical item are translated identically, one error may account for the need of edits in many sentences. For example, the expression *show everything about all X* was repeated in 20 of the 51 sentences included. In Danish, the noun form was incorrect (for example, *vis alt om alle maleri* instead of *vis alt om alle malerier* 'show everything about all paintings'), and this error accounts for 20 of the 31 cases where a sentence required editing in Danish. On the lexical side, one incorrect Norwegian noun (*oljemaling* 'oil painting (activity)' instead of *oljemaleri* 'oil painting (object)') occurred in seven sentences. In five cases, it was the only edit needed in the sentence.

To get a clearer picture of how often edits could be traced back to certain recurring cases, and what types of cases were involved, they were categorized according to the grammatical structure, rather than sentence. Altogether 20 different structures were used in the sentences, and one sentence could contain multiple structures – for example, '(NAME OF THE PAINTING) was painted on (MEDIUM) by (ARTIST) in (YEAR)'. Table 10 then shows the number of structures that had been corrected in each language.

The type of errors varied somewhat across languages. Noun form errors were found in Danish and Norwegian, while most French errors involved word order in different interrogative structures. In German, the evaluator had generally performed multiple edits per sentence and structure. Some of these were clearly matters of preference, like changing the translation for 'oil painting' from 'Ölmalerei' into the synonymous 'Ölgemälde'. Individual lexical items were less commonly edited, although pronoun errors were a relatively

Table 10: Corrected structures in the cultural evaluation sample

Instances	Structure	Cat	Dan	Fin	Fre	Ger	Nor	Swe
2	**(object) is painted in (colour)**	1	1	0	0	1	1	0
10	**(object) was painted**	1	0	0	1	0	0	0
5	**(objects) at (museum)**	1	0	0	0	0	1	0
5	**(objects) that are in (colour)**	0	1	0	0	1	1	0
5	**(objects) that are on (material)**	0	1	0	0	1	1	0
10	**by (painter)**	0	0	0	0	0	0	0
5	**how many (objects) are there**	1	1	0	1	1	1	0
10	**in (year)**	1	0	0	0	1	0	0
3	**it is of size (measure) by (measure)**	1	1	0	0	1	0	0
8	**on (material)**	0	0	0	0	0	0	0
1	**show (object)**	0	0	0	0	0	0	0
1	**show everything about (object)**	0	0	0	1	0	0	0
20	**show everything about all (objects)**	0	1	0	1	0	0	0
5	**this (object) is displayed at (museum)**	0	0	0	0	1	0	0
1	**what are the colours of (object)**	0	0	0	0	0	1	0
1	**what is the material of (object)**	1	1	0	1	1	1	0
1	**what is the size of (object)**	0	1	0	1	1	1	0
1	**when was (object) painted**	0	0	1	0	0	0	0
1	**where is (object) displayed**	0	0	0	1	1	1	0
1	**who painted (object)**	1	0	0	1	0	0	0
	Total	**8**	**8**	**1**	**8**	**10**	**9**	**0**

common source of changes. For example, in both Danish and Norwegian the interrogative pronoun 'how' had been translated incorrectly.

Some expressions were handled well by GF across all languages, such as 'show (object)' and 'by (painter)'. The different types of queries, on the other hand, appeared difficult in nearly all languages, with Swedish being the only one where no queries required editing. For some structures, correctness varied across languages. For example, 'was painted' had incorrect verb forms in both Catalan and French, but was correctly rendered in all other languages.

Although the number of sentences edited was very high in some languages, one strength of the rule-based GF approach is that all occurrences of a specific error can be corrected by changing the GF grammar.

As noted previously, only one Swedish sentence had been edited, and comparing the sentence to others with the same structure, it appears the post-editor had misunderstood one of the cases referring to 'gold' as material rather than colour. Generally, all occurrences of a given structure were either correct or incorrect (as evidenced by the evaluator corrections). One exception was found in Finnish, where one of the museum names had an incorrectly rendered grammatical case whereas all other names were inflected correctly.

7.2 Discussion of the Cultural heritage results

Some clear grammatical and lexical issues were identified in the evaluation. However, the fact that some languages were only evaluated by one evaluator makes it difficult to know if some of the edits may have been a matter of preference. For example, in French, the GF translation for 'show everything about' was given as 'montre toute l'information sur'. The French editor deleted the word 'information' in most cases, but left it untouched in others. In the languages with multiple evaluators, it was possible to compare the edits made by different people to see if specific structures or lexical items were changed by both/all or only one. Some variation is naturally expected. In Catalan, for example, one of the evaluators accepted all the cases of '(paintings) that are in (color)' while the other two changed some occurrences and accepted others – not in the same sentences, however. The evaluators seem to have found this structure unnatural, although apparently not incorrect.

8 General issues in human evaluation

It is well known that human evaluations are subjective to some degree – the choices of best translation and the edits performed are to some extent affected by individual evaluators' preferences. The evaluators can, and do, sometimes disagree with each other on the best suggestion, although the results do show high agreement overall. Although they all were instructed to make only minimal corrections, different people may have different opinion on what minimal corrections are needed. It is also possible that this understanding of minimal changes as the MT quality improves. When there are relatively few things that absolutely require editing (clear grammatical errors or incorrect words), the editors may become more prone to notice words or expressions that do not match their preferences and feel those need to be corrected than if there were clear errors in the MT.

Some issues also relate to the automatic metrics, which are based on lexical comparisons between the machine translation suggestion and a reference translation created by a human translator/post-editor. As such, it should be noted that they are capable of only measuring similarity between the suggestion and reference, not the quality of the MT suggestion directly.

One issue is that while it can generally be assumed that a suggestion identical to the human version is correct, a different suggestion is not necessarily incorrect. Similarly as the evaluator preferences affect their choice of the best translation, some differences evident in the automatic evaluation may be related to preferences in word choices. The number or differences or edit operations also does not necessarily describe the effort necessary in editing. For example, punctuation – which was not present in the current GF translations – can be added quite fast.

9 Conclusions

We have presented the results for several use-cases implementing the Grammatical Framework (GF) as a text generation and translation tool. Both the automatically calculated metrics and human evaluations show that GF translations in the limited domains like the tourist phrasebook, ACE-in-GF and mathematics perform very well compared to other off-the-shelf MT systems. Given a choice between several systems, GF translations are more frequently selected as a correct translation by evaluators, and post-editing GF suggestions requires less effort than the suggestions from other systems – the very low TER scores for these samples also demonstrate this. Except for some cases like the mathematic clauses for Finnish, there were very few syntactic issues in the evaluated examples. This shows that the abstract syntax used by GF is of high quality and can be adapted into languages with very different features.

In the more open domain of patent translations the GF systems perform at least as well as Google Translate. Both GF implementations got the rank of 'Useful' from the patents experts.

During the evaluation process, we have collected valuable information from native speakers with which we have already improved the GF grammars. It should be noted that by correcting the issues found in one phase of the evaluation, we have been able to improve the quality of the subsequent phases.

References

[1] Laura Canedo, Norbert E. Fuchs, Kaarel Kaljurand, Maarit Koponen, Tobias Kuhn, Jussi Rautio, and Victor Ungureanu. D11.3 Evaluations of ACE-in-GF and of AceWiki-GF. Technical report, MOLTO project, May 2013. http://www.molto-project.eu/biblio/deliverable/d113-evaluations-ace-gf-and-acewiki-gf.

[2] Dana Dannélls, España-Bonet, Aarne Ranta, Ramona Enache, Mariana Damova, and Maria Mateva. D8.3 W8 Translation and retrieval system for museum object descriptions. Technical report, MOLTO project, April 2013. http://www.molto-project.eu/biblio/deliverable/multilingual-grammar-museum-object-descriptions-0.

[3] George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[4] Cristina España-Bonet, Ramona Enache, Shafqat Virk, Erzsébet Galgóczy, Meritxell Gonzàles, Aarne Ranta, and Lluís Màrquez. D5.3 WP5 final report: statistical and robust MT. Technical report, MOLTO project, May 2013. http://www.molto-project.eu/biblio/deliverable/wp5-final-report-statistical-and-robust-mt.

[5] Christian Federmann. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35, September 2012.

[6] Jesús Giménez and Lluís Màrquez. Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. *The Prague Bulletin of Mathematical Linguistics*, (94):77–86, 2010.

[7] Gregor Leusch, Nicola Ueffing, Hermann Ney, and Lehrstuhl Für Informatik. A novel string-to-string distance measure with applications to machine translation evaluation. In *In Proceedings of MT Summit IX*, pages 240–247, 2003.

[8] Sonja Nießen, Franz J. Och, Gregor Leusch, and Hermann Ney. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, 2000.

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[10] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231, 2006.

[11] David Vilar, Jia Xu, Luis Fernando D'Haro, and Hermann Ney. Error Analysis of Machine Translation Output. In *International Conference on Language Resources and Evaluation*, pages 697–702, Genoa, Italy, May 2006.

A Patent rankings

A.1 German

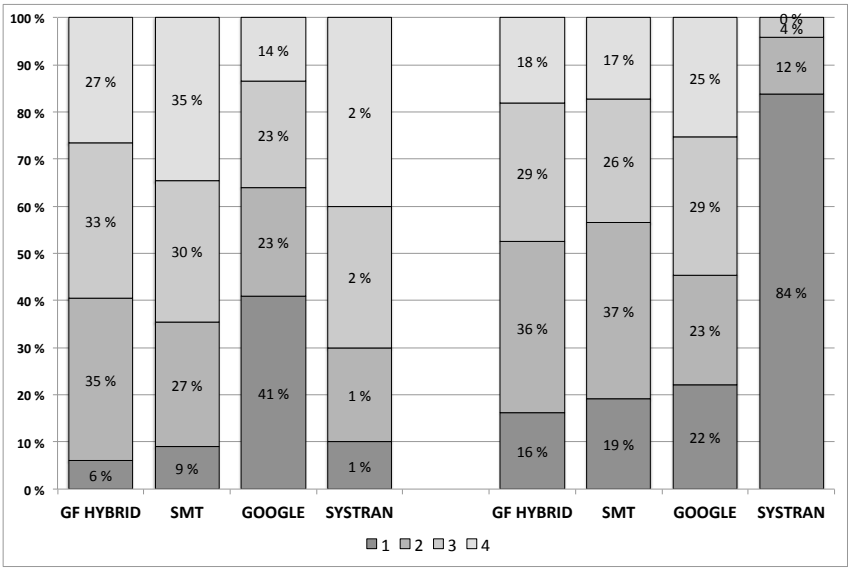

Figure 9: Rankings of the German EPOA61P by two evaluators

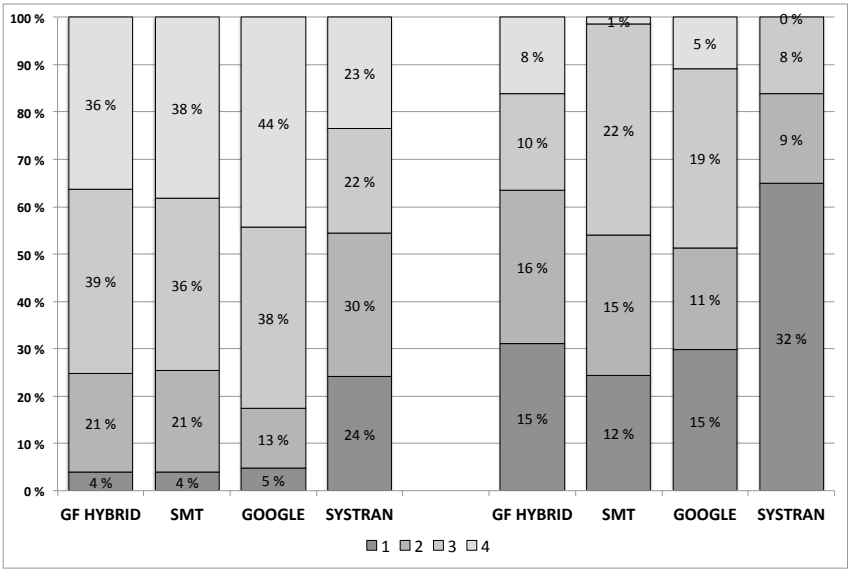

Figure 10: Rankings of the German PATSA61P by two evaluators

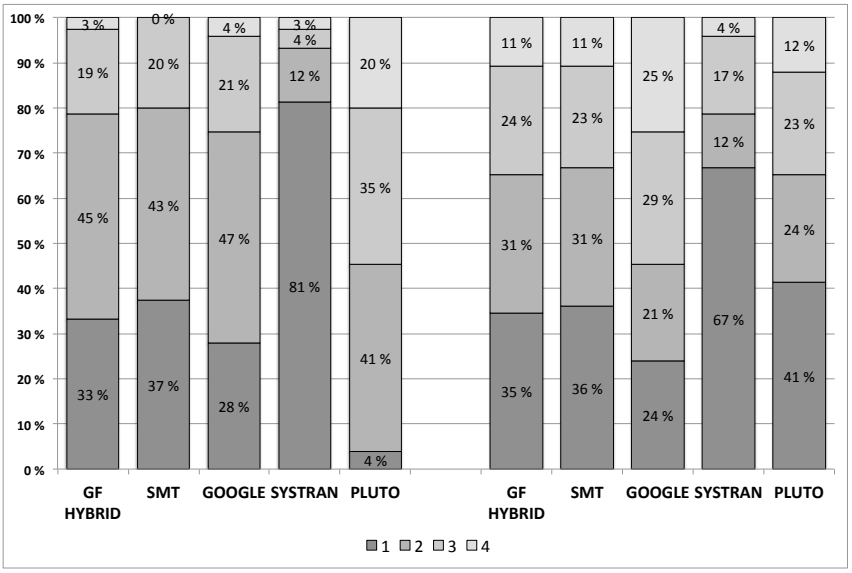

Figure 11: Rankings of the German USAPATS by two evaluators

A.2 French

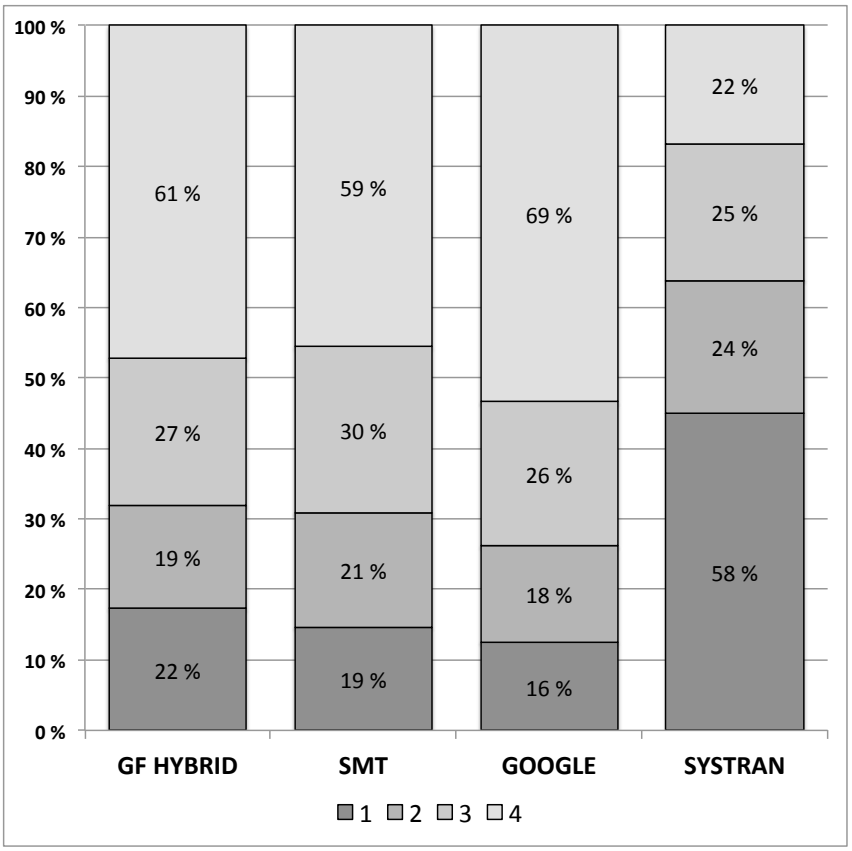

Figure 12: Ranking of the French EPOA61P

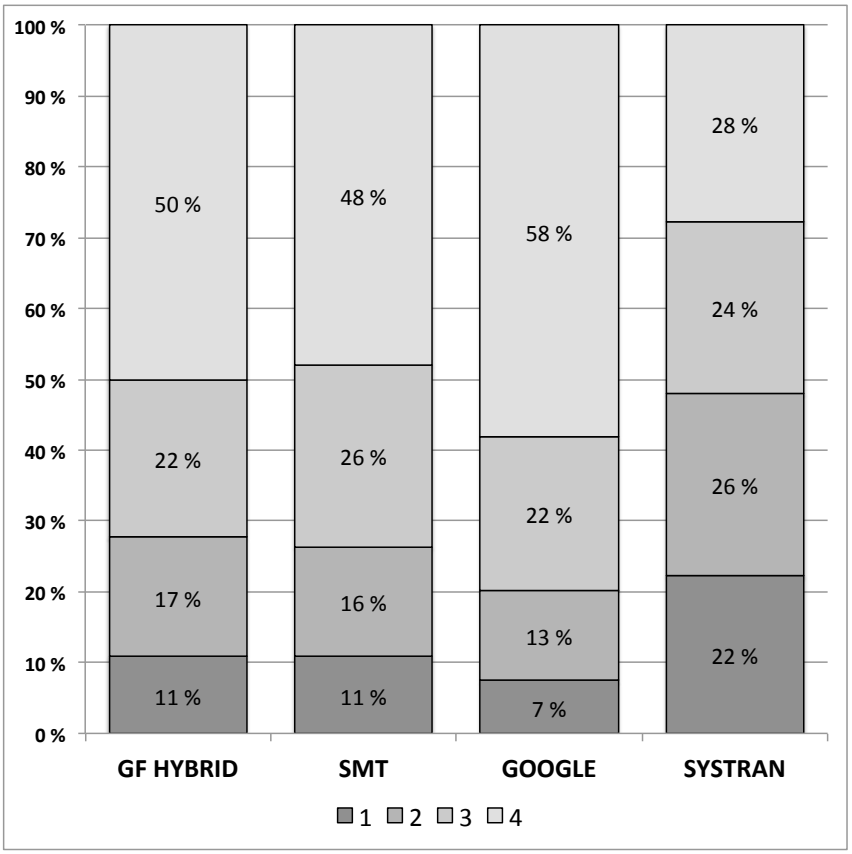

Figure 13: Ranking of the French PATSA61P

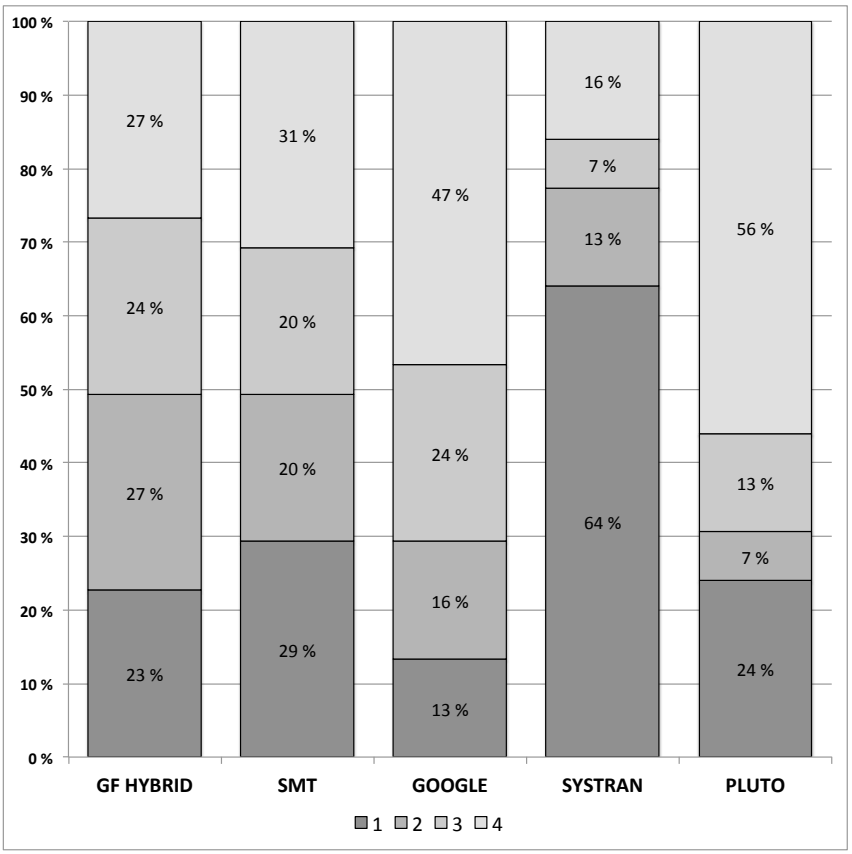

Figure 14: Ranking of the French USAPATS