Deliverable 5.2

Description and evaluation of the combination prototypes

Contract No.: FP7-ICT-247914
Project full title: MOLTO - Multilingual Online Translation
Deliverable: D5.2 Description and evaluation of the combination prototypes
Security (distribution level): Public, regular publication
Contractual date of delivery: M24
Actual date of delivery: March 15, 2012
Type: Regular Publication
Status & version: Draft v1.0
Author(s): Cristina España-Bonet, Lluís Màrquez, Ramona Enache, Aarne Ranta
Task responsible: UPC
Other contributors: UGOT

ABSTRACT
This document is the second deliverable corresponding to WP5, *Statistical and Robust Translation*. It is intended to be a description of the translation prototypes applied on the patents case study, WP7. These systems are able to exploit the high coverage of statistical translators and the high precision of GF to deal with specific issues of the language.

Contents

1 Introduction

This document is the second deliverable corresponding to WP5, *Statistical and Robust Translation*, and it is intended to be a description of the translation prototypes. The work here is tightly related to WP7, *Case Study: Patents*. The engines of the translation systems are built within WP5, and they are specialised and integrated into the prototype of the patents use case.

According to the project's description of work three kinds of translators are developed within the workpackage: two individual systems and several hybrid systems. One of the individual systems is a state-of-the-art phrase-based SMT system specialised for the patents domain, whereas the second one corresponds to a GF translator which, although follows the philosophy of the other WPs of the project, had to be adapted to a wider domain as it is patent translation with all the corresponding challenges it represents.

For the hybridisation prototypes two main approaches are implemented, what we call *hard integration* and *soft integration*. In the hard integration approach the GF partial output is fixed in a regular SMT decoding. In the soft integration approaches one of the two individual systems leads the translation and the other one complements.

The remaining of the document is as follows. Section 2 relates this work to current research approaches of the topics involved. Section 3 describes the domain of application of the work and the pre-processing necessary due to the characteristics of the domain. Afterwards, we introduce the SMT system (Section 4), the GF-based system (Section 5) and the combination prototypes (Section 6). In the three sections there is both the description of the system and its evaluation. Finally, we summarise and draw the conclusions according to the presented experiments and point out some improvements to the current systems.

2 WP5 work in context

This work tackles two topics which are lately attracting the attention of researchers, patent translation and hybrid translators.

2.1 Hybrid translation systems

The predominant core of machine translation (MT) systems has been changing through the years. From the very beginnings in the 50s where only dictionary-based MT systems existed, the technology evolved towards rule-based systems (RBMT). Later in the 90s the everyday more powerful computers allowed to develop empirical translation systems. Recently a type of empirical system, the statistical one (SMT), has become a widely used standard for translation. At this point the two main paradigms, RBMT and SMT, coexist with their strengths and weaknesses. Luckily these strengths and weaknesses are complementary and current efforts are being made to hybridise both of them and develop new technologies. A classification and description of hybrid translation can be found in [25].

In general RBMT provides high precision, due to an analysis of the text, but has limited coverage and a considerable amount of effort and linguistic knowledge is required in order to build such a system. On the other hand, SMT can achieve a huge coverage and is good at lexical selection and fluency but has problems in building structurally and grammatically correct translations.

Hybrid MT (HMT) is an emerging and challenging area of machine translation, which aims at combining the known techniques into systems that retain the best features of their components, and reduce the disadvantages displayed by each of the methods when used individually. Besides system combination strategies, hybrid models are designed so that there is one leading translation system assisted or complemented by other kinds of engines. This way the final translator benefits from the features of all the approaches. A family of models are based on SMT systems enriched with lexical information from RBMT [6, 3]. On the other side there are the models that start from the RBMT analysis and use SMT to complement it [11, 9, 8].

Our work on hybrid systems can be classified in the two families. On the one hand, SMT helps on the construction of the RBMT translator but, on the other hand, there is the final decoding step to integrate translations and complete those phrases untranslated by RBMT. With respect to the engines, a grammar-based translator is developed to assure grammatically correct translations. We extend GF (Grammatical Framework, [19]) and write a new grammar for patent translation. The SMT system that complements the RBMT is based on Moses [12].

The hybrids are specifically designed to deal with the translation of patents. The language of patents follows a formal style adequate to be analysed with a grammar, but at the same time uses a rich and particular vocabulary adequate to be gathered statistically.

2.2 Patent translation

The high number of patents being registered and the necessity for these patents to be translated into several languages are the reason so that important efforts are being made in the last years to automate its translation between various language pairs. Different methods have been used for this task, ranging from SMT [2, 7] to hybrid systems [4, 5]. Besides full systems, various components associated to patent translation are being studied separately [21, 22, 23].

In addition to this, the European project PLuTO (Patent Language Translations Online[1]) has also patent translation as one of the goals. PLuTO aims at making a substantial contribution by using a number of techniques that include hybrid systems combining example-based and hierarchical techniques.

[1] http://www.pluto-patenttranslation.eu/

3 Patents Domain

3.1 Relation with the patents case study, WP7

Patents have been chosen for the opening of the system to non-restricted language. This election has two main reasons. First, the language of patents, although having a larger amount of vocabulary and richness of grammatical structure than the mathematical exercises and the description of museum objects, still uses a formal style that can be interpreted by a grammar. And second, there is nowadays a growing interest for patents translation. The high and increasing number of registered patents has created a huge multilingual database of patents distributed all over the world. So, there is an actual need for building systems able to access, search and translate patents, in order to make these data available to a large community. Hence, the translation of patents text seems a natural scenario to test the techniques developed in this workpackage.

A patent is an official document granting a right. Besides the terms of the patent itself, it also contains information about its publication, authorship and classification for example. Being an official document, the structure giving the terms of the patent is quite fixed. Every patent has a title, a description, an abstract with the most relevant information and a series of claims. Here we want to deal with the translation of abstracts and claims.

A claim is a single (possibly very long) sentence composed mainly of two parts: an introductory phrase and the body of the claim, usually linked by a conjunction. It is in the body of the claim where there is the specific legal description of the exact invention. Therefore, claims are written in a lawyerish style and use a very specific vocabulary of the domain of the patent. All these features must be taken into account in order to define a corpus to be used for SMT, and to build the in-domain grammar.

3.2 Corpus

MOLTO works with European patents and the task is restricted to English, French and German. A first domain of application includes biomedical and pharmaceutical patents. We select patents with IPC (International Patent Classification) code A61P, corresponding to "Specific therapeutic activity of chemical compounds or medical preparations".

A parallel corpus in the three languages has been gathered from the corpus of patents given for the CLEF-IP track in the CLEF 2010 Conference[2] as described in deliverable 5.1. These data are an extract of the MAREC corpus, containing over 2.6 million patent documents pertaining to 1.3 million patents from the European Patent Office[3] (EPO). Our parallel corpus is a subset with those patents with translated claims and abstracts into the three languages. From this first subset we selected those patents that deal with the appropriate domain.

[2]http://clef2010.org/
[3]http://www.epo.org/

SET	Segments	EN tok	DE tok	FR tok
Training	279,282	7,954,491	7,346,319	8,906,379
Development	993	29,253	26,796	33,825
Test	1,008	31,239	28,225	35,263

Table 1: Numbers for the patents aligned corpus in English (EN), German (DE) and French (FR).

The final corpus built this way covers 56,000 patents out of the 1.3 million. That corresponds to 279,282 aligned parallel fragments as it can be seen in Table 1. A fragment is the minimum segment aligned in the three languages, so, it is shorter than a claim and, consequently, shorter than a sentence. Two small sets for development and test purposes have also been selected with the same restrictions: 993 fragments for development and 1008 for test.

Besides, the European Patent Office (EPO) recently provided us with a new corpus. Although it is composed of 1.7M fragments aligned by language pairs, less than 900 fragments (847 for German-English, 858 for French-English and 831 for French-German) correspond to the domain we tackle. This small dataset is going to be used for test purposes.

Notice that Table 1 shows the number of tokens for each of the languages, the smallest unit in a translation system. The selected corpus uses specific vocabulary of chemistry plenty of names of compounds, which have a particular structure that cannot be properly analysed with standard NLP tools. Therefore, the tokenisation process has been slightly modified. Next section introduces the linguistic processors used.

3.3 Linguistic processors

The detection and correct tokenisation of chemical compounds has been shown to be crucial in the performance of translators (see Section 4 for the analysis). A regular tokeniser would for example split the compound:

cis-4-cyano-4-(3-(cylopentyloxy)-4-methoxyphenyl)cyclohexane-1-carboxylic

into 9 tokens

cis-4-cyano-4-, (, 3-, (, cylopentyloxy,), -4-methoxyphenyl,), cyclohexane-1-carboxylic,

when using standard tokenisation rules. Consequently, each of the tokens would be translated as an independent word. To deal with this peculiarity of the domain, we developed a pipeline to detect, tokenise and translate compounds.

3.3.1 Compound recogniser and tokeniser

As a first approximation we devise a recogniser and tokeniser based on affix detection. A list with approximately 150 affixes has been compiled (142 elements for English and

German, and 148 for French) and it is used to select the candidate tokens to be a compound from the corpus. The list includes prefixes such as *Meth-, Eth-, Prop-, Pentadec-, imido-, selenocarboxy-, hydroxy-, Propion-, Arachid-...* and suffixes such as *-ol, -one, -al, -aldehyde, -oic, -oate, -oxy, -sulfonic, -nitrile, -amine* or *-isocyanide*.

The candidates selected this way are matched against a dictionary and those without a match are considered to be compounds and do not get an internal tokenisation. 103,272 compounds are found with this procedure within the training corpus defined in Section 3.2.

However, this list of compounds contains some noise. Examples of noise are in this context proper names with the defined affixes (*Hôpital*), words that do not appear in the dictionary (*extracorporeal*) or simply typos (*comparoate*). The amount of noise is considerable, but extra words do not in general imply a wrong tokenisation. So, the method works better as a (non-)tokeniser than as a compound detector and it bets for high recall instead of precision. Notice also that multiword compounds such as *Potassium bromide* cannot be detected with this methodology, but again, there is no negative effect on tokenisation. This pre-process is applied to the parallel corpus before training the SMT system.

Given the power of GF, one can also build a simple grammar for translating compounds. What makes the difference between this rule-based approach and a mere translation of each word in the compound is that in this case the possible reordering of the words is already defined by the grammar. So, functional words like *acid*, *ester* or *aldehyde* swap its position with the radical words whenever necessary. As explained in the Conclusions this is one of the improvements we plan to add to our systems.

3.3.2 Part-of-speech tagger, lemmatiser and named entity recogniser

Part-of-speech (PoS) tagging and lemmatisation are not used in the base statistical system but are necessary in the lexicon building of the patents grammar. `GENIA` [27], a linguistic processor prepared specially to process texts from the biomedical domain, is used for both purposes.

In addition to this we use named entity recognition as a pre-process for the translator. Named entities are marked in the text and are not translated by GF, but translated independently and substituted afterwards. In the biomedical domain we checked that a simple heuristic works as well tagging proper names as a state-of-the-art tagger not specifically trained (we used the Stanford PoS tagger [26]). In our experiments, we consider to be proper names the words starting with a capital letter (after lowercasing the beginning of the sentences), and the words containing numbers or special characters inside. A group of proper names is further on merged into only one proper name for simplicity, as they are proper names separated with hyphens, and proper names followed by a proper name between parentheses. This simple methodology lead to 100% precision and recall for the first 200 fragments in the training corpus of Section 3.2, where the proper names were manually annotated and the output was compared to that of the named entity recogniser. In this case 176 proper names were properly classified and replaced with a place holder name.

	DE2EN			EN2DE		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1−WER	0.52	0.64	**0.72**	0.42	0.51	**0.69**
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**

Table 2: Automatic evaluation using a set of lexical metrics of the in-domain SMT system for the English-German language pair. Results of two state-of-the-art systems, Bing and Google, are showed for comparison.

4 SMT system

In Statistical Machine Translation (SMT) and within the log-linear model [15], the best translation $\hat{e}$ for a given source sentence f is the most probable one, and the probability is expressed as a weighted sum of different elements:

$$T(f) = \hat{e} = \text{argmax}_{\text{e}} \sum_{m} \lambda_m h_m(f, e)\,. \tag{1}$$

In the standard most simple form, one considers 8 components being $h_m(f, e)$ log-probabilities of: the language model $P(e)$, the generative and discriminative lexical translation probabilities $lex(f|e)$ and $lex(e|f)$ respectively, the generative and discriminative translation models $P(f|e)$ and $P(e|f)$, the distortion model $P_d(e, f)$, and the phrase and word penalties, $ph(e)$ and $w(e)$.

The λ weights, which account for the relative importance of each feature in the log-linear probabilistic model, are commonly estimated by optimising the translation performance on a development set. For this optimisation one can use Minimum Error Rate Training (MERT) [14] where BLEU [17] is the reference score.

In our experiments, we build a state-of-the-art phrase-based SMT system trained on the biomedical domain with the corpus described in Section 3.2. Its development has been done using standard freely available software. A 5-gram language model is estimated using interpolated Kneser-Ney discounting with `SRILM` [24]. Word alignment is done with `GIZA++` [16] and both phrase extraction and decoding are done with the `Moses` package [13, 12]. The optimisation of the weights of the model is trained with MERT against the BLEU evaluation metric. Our model considers the language model, direct and inverse phrase probabilities, direct and inverse lexical probabilities, phrase and word penalties, and a non-lexicalised reordering.

Table 2 shows a first evaluation of this system (Domain) using a variety of lexical metrics. This set of metrics is a subset of the metrics available in the `Asiya` evaluation

	FR2EN			EN2FR		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1−WER	0.54	0.66	**0.78**	0.57	0.63	**0.73**
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**

Table 3: As in Table 2 for the English-French language pair.

	DE2FR			FR2DE		
METRIC	**Bing**	**Google**	**Domain**	**Bing**	**Google**	**Domain**
1−WER	0.42	0.52	**0.76**	0.30	0.43	**0.65**
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**

Table 4: As in Table 2 for the French-German language pair.

package [10]. We specifically select this set of metrics because all of them are available for the three languages: English, German and French. Together with our in-domain system we show the same evaluation for two public SMT systems for general translation: Bing[4] and Google[5]. These systems can be considered the state-of-the-art of an SMT open domain translator.

In general, our in-domain trained system performs significantly better than the two general purpose ones mainly because of two reasons. First, it has been trained on the specific domain and second, the tokenisation tools have been specifically developed to deal with chemical compounds. The concrete values can be read in Tables 2, 3 and 4 for the language pairs English-German, English-French and French-German respectively. As a general trend common to all systems, translations into German are the most difficult, and those involving English are easier due to the nature of the languages. French and German show richer morphologies which increases the complexity of the translation. This is therefore independent of the domain.

[4]http://www.microsofttranslator.com/

[5]http://translate.google.com

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 0.60 : 1 .
Google	The **method** of claim 23 , wherein the molar ratio of arginine to ibuprofen 0.60 : 1 **is** .
Bing	Use of claim 23 , wherein the molar ratio of arginine to ibuprofen is 0.60 : 1 .

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**
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

Table 5: Examples of wrong German-to-English translations in SMT systems. This kind of errors are not produced by the GF grammar for translating compounds.

Even though the Domain system shows a good performance among SMT systems, some of the observed translation errors would not be produced by a rule-based system, which, on the other hand, would probably produce different ones. Table 5 displays two translations from German into English where this is made evident. In the first one, systems are not able to capture the different order in the verb position, although the translation is adequate lexically. The second sentence is an example of the importance of the chemical names. Google, for instance, tokenises the compound by the punctuation. Some of the tokens are then translated, but the full compound is not recovered. Bing and Domain do not tokenise the compound, but according to the results, the word does not appear in the training corpus and has not been translated. These kinds of errors can be easily alleviated by the GF grammar and are a motivation to combine GF and SMT for the translation of patents.

5 GF system

GF is a type-theoretical grammar formalism, mainly used for multilingual natural language applications. Grammars in GF are represented as a pair of an *abstract syntax* –an interlingua that captures the semantics of the grammar on a language-independent level, and a number of *concrete syntaxes* –representing target languages. There are also two main operations defined, *parsing* text to an abstract syntax tree and *linearising* trees into raw text.

The GF resource library [18] is the most comprehensive grammar for dealing with natural languages, as it features an abstract syntax which implements the basic syntactic operations such as predication and complementation, and 20 concrete syntax grammars corresponding to natural languages. This layered representation makes it possible to regard

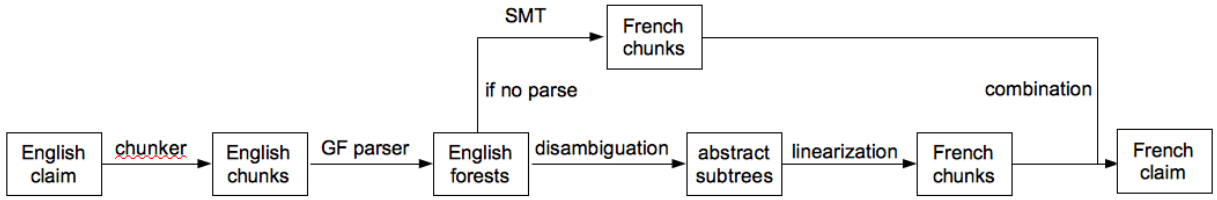

Figure 1: Architecture of the GF translation system.

multilingual GF grammars as a RBMT system, where translation is possible between any pair of languages for which a concrete syntax exists. However, the translation system thus defined is first limited by the fixed lexicon defined in the grammar, and secondly by the syntactic constructions that it covers. For this reason, GF grammars have a difficult task in parsing free text. There is some recent work on parsing the Penn Treebank with the GF resource grammar for English [1], whereas the current work on patent translation is the first attempt to use GF for parsing un-annotated free text.

The extension of GF to a new domain implies the construction of a specialised grammar that expands the general resource grammar. Since in our case of application we are far from a close and limited domain, some probabilistic components are also necessary. The general architecture is illustrated by Figure 1. A GF grammar-based system alone cannot parse most patent sentences. Consequently, the current translation system aims at using GF for translating patent chunks, and assemble the results in a later phase.

As explained in Section 3.3, claims are tagged with PoS with Genia as a pre-process. From the PoS-tagged words only the ones labelled as nouns, adjectives, verbs and adverbs are kept, since the GF library already has an extensive list of functional parts of speech such as prepositions and conjunctions. We use the extensive GF English lexicon[6] as a lemmatiser for the PoS-tagged words, so that one can build their correspondent abstract syntax entry. Moreover, all the inflection forms of a given word are obtained from the same resource.

This process is made online. For every sentence to translate, the lexicon is enlarged with the corresponding vocabulary. The French version of the lexicon is built by translating the individual entries from the English lexicon (all inflection forms) with the SMT individual system trained on the patent corpus. The French translations are lemmatised with an extensive GF French lexicon, based on the large morphological lexicon Morphalou [20] in order to get their inflection table. The part-of-speech is assumed to be the same as in the English counterpart.

When this procedure is applied on the test set, the part-of-speech tagger is able to find 2,013 lexicon entries. However, due to part-of-speech mismatching or to the fact that a given word was not found in the SMT lexical table, 43.81% of the entries could not be

[6]The GF English lexicon is based on the Oxford Advanced Learner's Dictionary, and contains around 50,000 English words.

translated to French.

In order to increase the coverage of the final GF translation, the grammar is adapted to deal with chunks instead of with full sentences. So, the source text is chunked into noun phrases (NP), adjective phrases (AP), adverbial and prepositional phrases (PP), relative pronouns (RP) and verb phrases (VP). Other kinds are ignored.

Some technical details have to be taken into account in order to build the patents grammar for chunks. Whereas NPs can be translated directly, a VP, RP or AP needs to have an NP to agree with, otherwise the GF grammar cannot know which linearisation form to choose. For NP and PP which can be translated independently, a mapping into corresponding GF categories is defined, whereas for VP, RP and AP, their GF mapping requires an NP in order to build their correspondent linearisation. If the required NP is not found, the chunk is sent to the SMT. Also, the VP category from the English and French GF resource grammars is implemented as a discontinuous category, so that it can handle discontinuous constituents in English and clitics in French. The patent grammar uses a category built on top of VP, which represents the flattened version of a VP, with all the constituents combined.

Because the syntactical structure of chunks is important in this case, a post-processing step is needed. This is meant to ensure that the PoS-tagging is consistent and that certain aspects captured in the grammar can be properly reflected in the claims. One can see the importance of this step with an example.

Ex1 *The use of claim 1 , wherein said use is intramuscular .*

In the previous example, "said", a frequent used word in patent claims, acts as a definite article, whereas Genia tags it as a verb and therefore is it not merged with the following noun into a noun phrase. Moreover, the relative pronoun "wherein" is labelled as an adverb or noun phrase. The post-processing process updates the tags of certain entries and the tag of the following word, when needed.

Table 6 shows how the original tagging from Genia is converted into the correct GF parse chunks: *the use* (NP), *of claim 1* (PP), *wherein* (RP), *said use* (NP), *is intramuscular* (VP). As one can notice, chunks are merged when needed, like for the PP *of claim 1*, where the preposition was merged with the NP into a single chunk. The same goes for the VP chunk, as it is aimed to combine two-placed verbs or copulas with their objects before parsing.

GF parses the corresponding English chunks to obtain a forest of abstract syntax trees. In order to disambiguate among the possible options, all of them are linearised, looked up in the French corpus and the most frequent linearisation is kept as the best translation.

The translation sequence is done from left to right, so that the last-occurring NP is retained, and is used to make the agreement with VP, RP or AP. If no such NP can be found, or if the GF grammar is not capable to parse the one indicated by the chunker, the current chunk is passed to the SMT. In the working example, this in not necessary, and GF grammar alone obtains a translation for the full sentence:

1. *the use* → "l' utilisation" (NP)

Word	PoS Genia	Chunk Genia	PoS Final	Chunk Final
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
.	.	0	.	O

Table 6: Chunk detection for the example sentence Ex1.

2. *of claim 1* → "selon la revendication 1" (PP)

3. *wherein* → "dans laquelle" (RP agreeing with *"l' utilisation"'*)

4. *said use* → "ladite utilisation" (NP)

5. *is intramuscular* → "est intramusculaire" (VP agreeing with *"ladite utilisation"*)

Finally, chunks are combined together with the punctuation marks, other non-included elements and untranslated chunks in the same order as in the source language. Due to the nature of this system results are analysed together with those of the hybrid systems in the following section.

6 Hybrid systems

As it has just been seen, the grammar-based translator already makes use of the SMT system trained on patents to translate the GF English lexicon. Although it already uses of hybridisation techniques, we consider this first approximation as a baseline for the more advanced hybrid systems. The reason is that even the vocabulary is disambiguated towards the biomedical domain thanks to the hybridisation, still there are non-parseable chunks with unknown vocabulary in the lexicon that cannot be translated using the grammar. That is to say, the system is not able to translate robustly a whole test set. The percentage of sentences that can be completely translated from beginning to end by GF is of a 6.9%.

In the description of work three baseline systems were proposed: the individual SMT system, the individual GF system, and a naïve cascade combination where a sentence is translated by GF whenever is possible and by SMT otherwise. After the development of

	GF	SMT
NP	2,366 (14.9%)	2,199 (13.8%)
VP	275 (1.7%)	1,302 (8.2%)
AP	1,960 (12.3%)	1,935 (12.2%)
RP	648 (4.1%)	86 (0.5%)
Other	–	5,099 (32.0%)
Total	*5,301 (*33.3%*)*	*10,621 (*66.7%*)*

Table 7: Number and percentage of individual chunks translated by the HI system.

the systems we have defined two different baseline systems: the SMT system and the first hybrid GF translator. The latter is already a hybrid baseline and the naïve combination of both does not provide new information since more than 93% of the sentences would be translated by SMT.

To gain robustness in the final system the output of the GF translator is used as *a priori* information for a higher level SMT system. An SMT system trained in the same way as the SMT baseline is fed with these GF phrases and it is the way the phrases interact with the SMT ones what defines two families of hybrid models:

Hard Integration (HI): Phrases with GF translation are forced to be translated this way. The system can reorder the chunks and translates the untranslated chunks, but there is no interaction between GF and pure SMT phrases.

Soft Integration (SI): Phrases with GF translation are included in the translation table with a certain probability so that the phrases coming from the two systems interact. Probabilities in the SMT system are estimated from frequency counts in the usual way; the probabilities in the GF system are for the moment assigned to compete with the frequentists.

Under this perspective, the GF translator can be seen as a Soft Integration model led by a GF translator instead of being led by the SMT. The patents grammar parses the source sentence and complements its translations by the ones given by the SMT translation of the lexicon. These three integrations are evaluated on the patents test set both automatic and manually.

The chunking step is common in all the cases and after the pre-process, the test set is divided in 15,922 chunks. From these chunks a 33.3% can be translated using the GF patents grammar, and the remaining 66.7% must be passed to the SMT system. Table 7 shows the concrete percentages for every kind of chunk. Notice that GF only is designed to deal with the four most frequent types of chunks, and punctuation and conjunctions for example are ignored by GF. For these majority categories, GF can handle half of NP and AP, almost all RP but only a 17.4% of VP.

	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

Table 8: Automatic evaluation of the baselines and hybrid systems.

There are several reasons why GF cannot translate the chunks. In a 18.3% of the cases the chunks could not be parsed by the GF English grammar. When parsed, a 15.5% of the chunks could not be translated due to missing words in the bilingual lexicon and to a lesser extent a 1.1% could not be translated because of the missing information about agreement. A 31.3% of the chunks are labelled as *Other* (punctuation marks, item markers, etc.) and ignored by GF.

Splitting the sentences in chunks proved to be crucial for the final translation. A 84.7% of the fragments to be translated contained at least one chunk that could not be parsed by the English grammar, and even more, a 93.1% of the fragments contained at least one chunk that could not be translated. So, the coverage of a GF translation at sentence level would be of only a 6.9%. At chunk level the coverage increases up to a 33.3%.

Still this limited coverage cannot compete with that of a statistical system. Table 8 reports an automatic evaluation using several lexical metrics for both GF and SMT individual systems (top rows). For all the metrics the SMT system beats the GF one in a significant way. This is mainly due to the coverage, SMT is able to translate the whole sentence which is not the case of GF. However, GF is able to deal with some grammatical issues that cannot be recovered statistically. The most evident example is agreement in gender and number. Contrary to English, French adjectives and nouns agree in gender and number and relative pronouns agree with their relative. This is taken into account by construction in GF so that mistaken SMT translations such as "le médicament séparée" is correctly translated as "le médicament séparé" (*the separate medicament*) or "composition pharmaceutique selon la revendication 1, dans lequel" is correctly translated as "composition pharmaceutique selon la revendication 1, dans lequelle" (*the pharmaceutical composition of claim 1, wherein*).

These are minor details from the point of view of the lexical evaluation metrics however, they make a difference to the reader. Although in few occasions the understanding of the sentence is compromised because of the lack of agreement, the fluency of the output is harmed.

Therefore we incorporate these well-formed translations into the SMT system. A hard integration of the translations does not allow them to interact. GF translations are always used and the statistical decoder reorders them and completes the translation with its own

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

Table 9: Manual evaluation of the 23 different sentences from a random subset of 100 sentences.

phrase table. This system is named HI in Table 8. Results are below those of the SMT system because the system is being forced to use the high quality translations together with translations of elements not considered. Just to give an example, GF will highly benefit from incorporating a grammar to deal with compounds and numbers. Currently these elements typical of the domain are not specifically approached.

A softer integration of the translations is done by the family of systems denoted by SI in Table 8. In this case, GF translations are given a probability which ranges from null to one. The probability is given to the chunk translation as a whole, so when competing with SMT translations that have four translation probabilities (phrase-to-phrase and word-to-word) the probability mass is divided among them and a value of one does not imply a sure translation. We show in the bottom rows of Table 8 just three different values for the probabilities: 0, 0.5 and 1. Relative probabilities between the systems result not to be as important as the fact of allowing the interaction.

The combination of all the phrases improves the translations according to all the lexical metrics considered. There is an increment of 0.42 points of BLEU, 0.30 of TER and 0.46 of ULC, an uniform linear combination of 13 variants of the metrics considered. Improvements are moderate because of two reasons. First, SMT translations are already good for a start. Second, the amount of issues that GF handles are limited to be reflected on automatic metrics.

In order to check this point and analyse the fluency of the outputs, we have conducted a manual evaluation of the translations. To do this, 100 sentences have been randomly selected and four evaluators have been asked to indicate the grammatically most correct translation between two options: the SMT translation and the SI0.5 hybrid translation (best hybrid system). The first thing to notice is that 77 sentences out of the 100 are identical in the two systems. The results for the remaining 23 can be seen in Table 9. The hybrid system is better than the SMT one according to the four evaluators, and the improvements come from discrepancies in gender, number and agreement. The SMT translations were preferred in the cases where the hybrid translation failed to translate certain words, so that the final claim has a visible hole –which makes it syntactically incorrect.

Figure 2 shows an example sentence where these features are observed. GF is doing

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**...

Figure 2: Example where GF translates with the correct gender of the adjective and the SMT completes the untraslated words.

the gender agreement between noun and adjective correctly ("séparée" vs. "séparé") but is not able to translate the full sentence ("at the same time as"). The two hybrid systems in this case are able to construct the correct translation which coincides with the reference.

7 Conclusions

We have developed and evaluated the first combination prototypes for the translation of patents within MOLTO. The systems exploit the high coverage of statistical translators and the high precision of GF to deal with specific issues of the language.

At this moment the grammar tackles agreement in gender, number and between chunks, and reordering within the chunks. Although the cases where these problems apply are not extremely numerous both manual and automatic evaluations consistently show their preference for the hybrid system in front of the two individual translators.

The coverage of the grammar can be extended in order to deal with more typical structures present in patent documents. The coverage of VP is particularly low because of the missing verbs from the French lexicon and the syntactically complex verb phrases –such as cascades of nested verbs, which are not handled by the patents grammar yet. Also, a grammar to translate compounds will be included as they are a significant part of the biomedical documents. Moreover, the grammar component can be extended to handle the ordering at sentence level besides of the reordering within the chunks. This is specially interesting to deal with languages like German where the structure of the sentence is different from the structure in English for example.

The previous improvements will increase the number of chunks that can be parsed by the grammar; in order to increase the percentage of translations it is also necessary to improve the lexicon building procedure. An obvious improvement would be a bilingual dictionary of idioms, so that the translation would not just map word-to-word, but also phrase-to-phrase.

Finally, we plan to implement another version of the hybrid system where GF grammars are applied at a later stage –after the English chunks are translated into French by the SMT system. The GF grammars will be used to to restore the agreement for chunks like VP, RP and AP, like before. The main difference is that due to an earlier use of SMT, one can capture idiomatic constructions better, and use GF just in the end for improving syntactic correctness.

We therefore do not consider the reported systems to be the definitive ones, we plan to enhance them in several aspects and also include a German grammar to complete the translation in the three languages with all the systems.

7.1 Dissemination

The work included in this report has been submitted to two refereed conferences. The SMT system and the work related to the pre-process of the data has been published in:

- **Patent translation within the MOLTO project**
  Cristina España-Bonet, Ramona Enache, Adam Slaski, Aarne Ranta, Lluís Màrquez and Meritxell Gonzàlez
  Proceedings of the 4th Workshop on Patent Translation, MT Summit XIII, Xiamen, China, September 23, 2011.

Whereas the latest work on hybridisation has been recently submitted as:

- **A Hybrid System for Patent Translation**
  Ramona Enache, Cristina España-Bonet, Aarne Ranta and Lluís Màrquez
  Submitted to 16th Annual Conference of the European Association for Machine Translation, EAMT 2012.

References

[1] Angelov, K. *The Mechanics of the Grammatical Framework*. PhD thesis, Chalmers University of Technology, Gothenburg, Sweden, 2011.

[2] Ceausu, A., Tinsley, J., Way, A., Zhang, J., and Sheridan, P. Experiments on Domain Adaptation for Patent Machine Translation in the PLuTO project. *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 2011)* (2011).

[3] Chen, Y., and Eisele, A. Hierarchical hybrid translation between english and german. In *Proceedings of the 14th Annual Conference of the European Association for Machine Translation* (5 2010), V. Hansen and F. Yvon, Eds., EAMT, EAMT, pp. 90–97.

[4] Ehara, T. Rule based machine translation combined with statistical post editor for japanese to english patent translation. *MT Summit XI Workshop on patent translation, 11 September 2007, Copenhagen, Denmark* (2007), 13–18.

[5] Ehara, T. Statistical Post-Editing of a Rule-Based Machine Translation System. *Proceedings of NTCIR-8 Workshop Meeting, June 15–18, 2010* (2010), 217—-220.

[6] Eisele, A., Federmann, C., Saint-Amand, H., Jellinghaus, M., Herrmann, T., and Chen, Y. Using moses to integrate multiple rule-based machine translation engines into a hybrid system. In *Proceedings of the Third Workshop on Statistical Machine Translation* (2008), StatMT '08, pp. 179–182.

[7] España-Bonet, C., Enache, R., Slaski, A., Ranta, A., Màrquez, L., and Gonzàlez, M. Patent translation within the molto project. In *Proceedings of the 4th Workshop on Patent Translation, MT Summit XIII* (Xiamen, China, sep 2011), pp. 70–78.

[8] España-Bonet, C., Labaka, G., Díaz de Ilarraza, A., Màrquez, L., and Sarasola, K. Hybrid machine translation guided by a rule-based system. In *Proceedings of the 13th Machine Translation Summit* (Xiamen, China, sep 2011), pp. 554–561.

[9] Federmann, C., Eisele, A., Chen, Y., Hunsicker, S., Xu, J., and Uszkoreit, H. Further experiments with shallow hybrid mt systems. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR* (July 2010), pp. 77–81.

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

[11] Habash, N., Dorr, B., and Monz, C. Symbolic-to-statistical hybridization: extending generation-heavy machine translation. *Machine Translation 23* (2009), 23–63.

[12] Koehn, P., Hoang, H., Mayne, A. B., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session* (Jun 2007), pp. 177–180.

[13] Koehn, P., Shen, W., Federico, M., Bertoldi, N., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Bojar, O., Zens, R., Constantin, A., Herbst, E., and Moran, C. Open Source Toolkit for Statistical Machine Translation. Tech. rep., Johns Hopkins University Summer Workshop. http://www.statmt.org/jhuws/, 2006.

[14] Och, F. J. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics* (Sapporo, Japan, July 6-7 2003).

[15] Och, F. J., and Ney, H. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (2002), pp. 295–302.

[16] Och, F. J., and Ney, H. A systematic comparison of various statistical alignment models. *Computational Linguistics 29*, 1 (2003), 19–51.

[17] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association of Computational Linguistics* (2002), pp. 311–318.

[18] Ranta, A. The GF resource grammar library. *Linguistic Issues in Language Technology 2*, 1 (2009).

[19] Ranta, A. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford, 2011. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

[20] Romary, L., Salmon-Alt, S., and Francopoulo, G. Standards going concrete: from lmf to morphalou. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries* (Stroudsburg, PA, USA, 2004), ElectricDict '04, Association for Computational Linguistics, pp. 22–28.

[21] Sheremetyeva, S. Natural language analysis of patent claims. In *Proceedings of the ACL-2003 workshop on Patent corpus processing - Volume 20* (Stroudsburg, PA, USA, 2003), PATENT '03, Association for Computational Linguistics, pp. 66–73.

[22] Sheremetyeva, S. Less, Easier and Quicker in Language Acquisition for Patent MT. *MT Summit X, Phuket, Thailand, September 16, 2005, Proceedings of Workshop on Patent Translation* (2005), 35–42.

[23] Sheremetyeva, S. On Extracting Multiword NP Terminology for MT. *EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation, ed. Lluís Màrquez and Harold Somers, 14-15 May 2009, Universitat Politècnica de Catalunya, Barcelona, Spain* (2009), 205–212.

[24] Stolcke, A. SRILM – An extensible language modeling toolkit. In *Proc. Intl. Conf. on Spoken Language Processing* (2002).

[25] Thurmair, G. Comparing different architectures of hybrid machine translation systems. In *Proc MT Summit XII* (2009).

[26] Toutanova, K., Klein, D., Manning, C., and Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL* (2003), p. 252–259.

[27] Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T., McNaught, J., Ananiadou, S., and Tsujii, J. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics.*, P. Bozanis and e. Houstis, E.N., Eds., vol. 3746. Springer Berlin Heidelberg, 2005, p. 382–392.