Deliverable 5.1
Description of the final collection of corpora

Contract No.: FP7-ICT-247914
Project full title: MOLTO - Multilingual Online Translation
Deliverable: D5.1 Description of the final collection of corpora
Security (distribution level): Public, regular publication
Contractual date of delivery: M18
Actual date of delivery: September 16, 2011
Type: Report
Status & version: v1.1
Author(s): Cristina España-Bonet, Meritxell Gonzàlez, Lluís Màrquez
Task responsible: UPC

ABSTRACT
The present document reports the corpora collection needed for the translation systems developed within the workpackage. First of all, it is introduced the framework and domain of application of the workpackage, with a special interest to the structure of patents. If follows a description of the methodology and content of the in-domain and out-of-domain corpora. Finally, we summarise the current status of the workpackage with relation to the data collection.

Contents

1 Introduction

This document is the first deliverable corresponding to WP5, *Statistical and Robust Translation*. It is intended to be a description of the final collection of corpora to be used within the workpackage.

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. Therefore, a part of the collection of data is that obtained for patents in WP7.

Empirical translation systems in general, and statistical machine translation systems in particular, need of a large amount of data. In these engines parallel corpora are used to train the translation model and a monolingual corpus is used to build a language model. For the languages tackled here (English, French and German) there are large corpora available to the machine translation community such as the European Parliament Corpus[1], the United Nations corpus[2] and several news sources (e.g. News Commentary parallel corpus[3]).

However, all of these common sources allow to gather a general purpose corpus with different characteristics from the patents domain of WP7. It is interesting to study the behaviour of the translation systems not only with this out-of-domain data, but also with a specific corpus of patents. Statistical systems show a better performance when trained on in-domain data, but the nature of the data is even more important for rule-based systems. In this case, one does not need a large corpus but a representative one. Rules or abstract syntax in our case are written by inspection of those structures that appear in patents. Therefore, the major task corresponding to this deliverable has been the collection of a specialised corpus build up with patents, in particular patents corresponding to the biomedical domain.

With this intention, we use the CLEF-IP data provided in the CLEF 2010 Conference[4]. The data is a mixture of European Patent Office[5] (EPO) patent applications and granted patents. Granted patents have the claims translated into English, French and German. To the day of publication of this deliverable, MOLTO has at its disposal only personal licenses for the usage of the corpus but the consortium is in negotiations with EPO since December 2010 in order to get a more general license.

The remaining part of this document is devoted to describe in detail both the in-domain and the out-of-domain corpora. But before this, Section 2 introduces the framework and domain of application of the workpackage. Afterwards, the in-domain corpus and the out-of-domain ones are described in Section 3 and Section 4 respectively. Finally, Section 5 summarises the current status of the workpackage with relation to the data collection.

[1] http://www.statmt.org/europarl/
[2] http://www.uncorpora.org/
[3] http://www.statmt.org/wmt11/
[4] http://www.ir-facility.org/clef-ip
[5] http://www.epo.org/

2 Statistical and Robust Translation

The purpose of this workpackage is to develop translation methods to enhance the quality and precision of grammar-based methods with the coverage and robustness of corpus-based ones. The focus is placed on techniques for combining Grammatical Framework (GF) and Statistical Machine Translation (SMT) systems.

The aim is the obtaintion of a hybrid translation system having both high quality and high coverage in the quasi-open domain of patents. Several variants which can be grouped in three families will be studied:

Hard integration Force fixed GF fragment translations within a SMT system.

Soft integration led by SMT Make available GF fragment translations to a SMT system.

Soft integration led by GF Complement with SMT options the GF translation structure.

GF [Ran11] is the main technology behind MOLTO. Its main feature is the notion of multilingual grammars, which describe several languages simultaneously by using a common representation, called abstract syntax. Because of the way the multilingual grammar is structured, it can also be used as a rule-based machine translation system between any pair of languages, for which a concrete syntax is provided. This way, meaning-preserving translation is automatically provided as a composition of parsing and generation via the abstract syntax, which works as a semantic interlingua.

In cases like the two restricted domains chosen by MOLTO, mathematical exercises (WP6) and description of museum objects (WP8), a grammar must be built in order to complement the general resource grammar and no corpus is necessary. The translation system is restricted to the language generated by the grammar –a controlled language with limited vocabulary and limited set of constructions.

By contrast, the language of the patents domain (WP7) is much broader. Even though one can construct a domain-specific grammar with the structures characteristic of the patent corpus, free in-domain text cannot be fully translated. A key issue, but not the only, is the limited lexicon. Since the vocabulary of patent claims is virtually unlimited, the lexicon for the patents grammar cannot be defined beforehand and it must be built at translation time. Just to give an example, the translation of the first 200 fragments in the training corpus (see Section 3.1) involves a lexicon of almost 700 entries.

For the SMT component, a corpus is a must. Our statistical system is a state-of-the-art phrase-based SMT system trained on the biomedical domain with the corpus described in Section 3.1. So, all the components of the hybrid systems need in a way or another an in-domain corpus, which, in our case of study, is the biomedical domain.

2.1 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.

2.2 In-domain data: patents

A patent is an official document granting a right. The file or files associated to every patent contain not only the terms of the patent itself, but also bibliographic data (i.e. publication, authorship and classification). Being an official document, the structure giving the terms of the patent is quite fixed. The documents are normalised to an XML format, in which the standardised fields include dates, countries, languages, references, person names, and companies as well as rich subject classifications. Every patent has a title, a description, an abstract with a short and general summary and a series of claims.

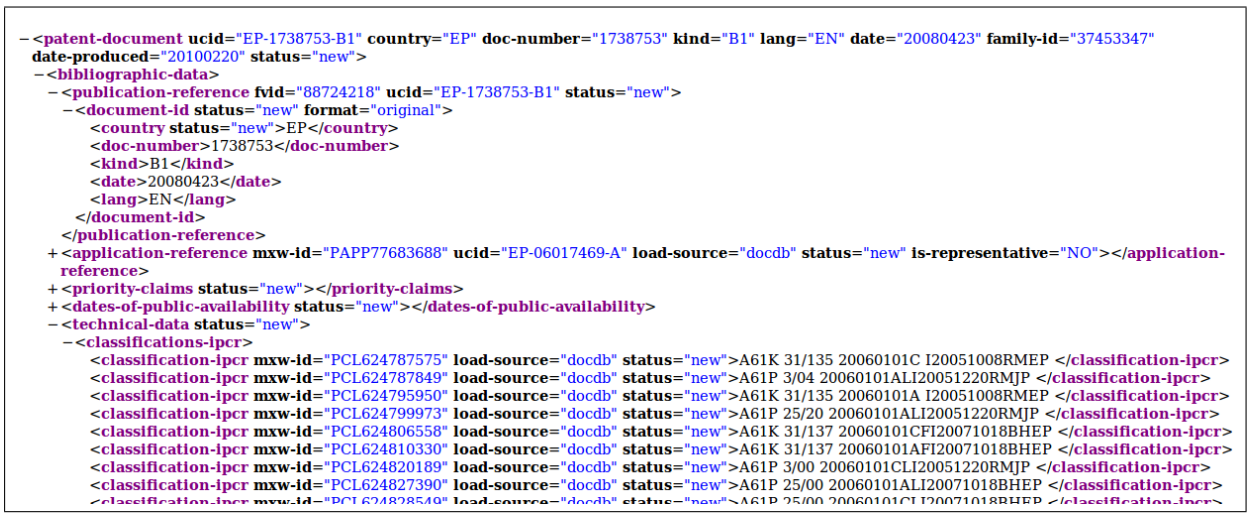

Figure 1: An extract of the bibliographic data of a patent document.

Figure 1 contains an excerpt of the bibliographic data from a patent document. This example shows the basic data supplied by all the documents: the two letter country code (*EP* –European in this case–), date (*20081423*) and language (*EN*) among others. In the specific corpus (Section 3.1), the documents from different countries and sources have

been normalised to a common XML format with a uniform patent numbering scheme and citation format.

The technical data is a relevant section of the document. It contains the list of IPC[6] codes assigned to the patent, i.e. the classification of the patent according to the different areas of technology to which the patent pertain. The IPC is arranged in a hierarchical structure of 8 sections, divided into 120 classes, 600 subclasses and 70,000 groups. The following are the title of the sections, the highest level of hierarchy:

A - Human Necessities

B - Performing Operations, Transporting

C - Chemistry, Metallurgy

D - Textiles, Paper

E - Fixed Constructions

F - Mechanical Engineering, Lighting, Heating, Weapons, Blasting

G - Physics

H - Electricity

The two digits following the section symbol correspond to the class, the second level of the hierarchy (*A61* is the symbol for the class "Medical or Veterinary Science and Hygiene"). Then, each class comprises one or more subclasses, and it is indicated with a capital letter following the class digits. Each subclass is broken down into groups (one- to three-digit number followed by a slash and 00) and subgroups (integer counting from 01 at the right side of the slash). A patent can be classified into more than one class, as seen in the bibliographic data in Figure 1.

The textual elements of a patent are the abstract, the description and the claims. Despite most of the patent documents contain abstracts, usually they do not provide descriptions[7]. Each of the three sections has a different purpose: the abstract gives the most relevant information of the invention, the description gives background information for understanding the invention and the series of claims constitute the legal scope of protection of the patent.

A claim is a single (possibly very long) sentence composed mainly of two parts: an introductory phrase and the body of the claim. As it has been said, 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. The following sentences illustrate such characteristic:

[6]International Patent Classification, http://www.wipo.int/classifications/ipc/en/

[7]Data source: http://www.ir-facility.org/prototypes/marec/statistics

- *The use according to claim 7, wherein said cancer diseases comprise bladder, lung, mamma, melanoma and prostate carcinomas.*
- *The pharmaceutical composition according to claim 1 or 2, wherein said platinum anticancer agent is selected from at least one of the complexes having structures of: **IMAGE**.*

Example 2: Lawyerish language (in blue) and specific vocabulary (in red) in the biomedical domain.

An excerpt of a patent description and the text of the patent claim can be seen in Figure 3. At the top of the figure there is a fragment of the description written in English. Next, the series of claims in the available languages are listed. Each claim consists of a sequence of texts which may contain figures, formulae or, as in this case, chemistry items.

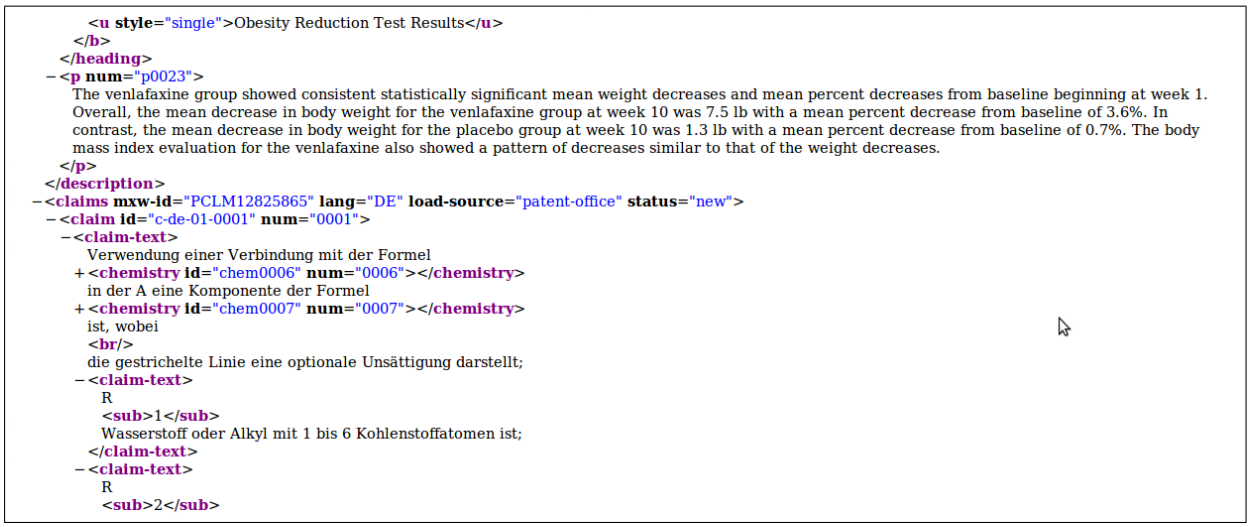

Figure 3: An extract of the description and claims sections of a patent document.

3 In-domain corpus

As an European project that aims to translate among the European languages, MOLTO works with European patents. The EPO is then a natural provider for the data. They register all the patent entries at least in their three official languages (European Patent Convention (EPC), Art.14) and therefore, our task is restricted to these three languages: English, French and German.

Despite the original language in which a patent is written, the specifications of the European patents are published in one of the official languages, and shall include a translation of the claims in the other two official languages. In consequence, the EPO is creating three

parallel corpora of claims in the three official languages and several bilingual corpora made up of descriptions in the European languages [TÏ1].

Up to now MOLTO lacks the EPO data although the consortium is in negotiations with the organisation. Alternatively, we have at our disposal a personal research license for the MAREC corpus, a larger corpus with a subset of EPO patents. The next section describes the characteristics of this corpus.

3.1 MAREC and CLEF-IP corpora

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[8]. These data are an extract of the MAREC corpus, containing over 2.6 million patent documents pertaining to 1.3 million patents from the EPO with some content in English, German and French, and extended by documents from the WIPO[9].

MAREC corpus

MAREC is a data collection over 19 million of European, US and Japanese patent applications and granted patents from 1976 to June 2008. In MAREC, the majority of the documents are written in English, German and French, and about half of the documents include full text. The documents follow a unified XML format normalised from sources of the European Patent Office, World Intellectual Property Organisation, United States Patent and Trademark Office and Japanese Patent Office (only applications). The standardised fields include dates, countries, languages, references, person names, and companies as well as rich subject classifications.

Not all documents in MAREC have all sections. Table 1 shows the distribution on abstracts and descriptions among the MAREC documents and Table 2 shows the same figures among languages. These data are provided by the IRF website[10].

Patent Office	**Documents**	**with at least one description**	**with at least one abstract**	**Abstracts**
EPO	*3,508,686*	*1,887,745*	*1,530,737*	*1,567,162*
WIPO	1,784,980	1,245,798	1,694,960	1,694,988
US	5,639,471	5,599,940	5,304,678	5,307,284
JP	8,453,560	0	8,453,560	8,453,560
Total	19,386,697	16,983,935	17,022,994	8,733,483

Table 1: Distribution on abstracts and descriptions among the MAREC documents.

[8] http://clef2010.org/

[9] World Intellectual Property Organization, http://www.wipo.int

[10] http://www.ir-facility.org/prototypes/marec

Abstracts

Patent Office	EN	FR	DE	Other
EPO	*1,371,278*	*107,398*	*383,644*	0
WIPO	1,694,139	1,578,844	191,351	10,236
US	5,304,678	0	0	0
JP	8,453,560	0	0	0
Total	16,823,655	1,686,242	574,995	1,0236

Descriptions

Patent Office	EN	FR	DE	Other
EPO	*1,277,752*	*144,388*	*465,604*	0
WIPO	979,986	60,631	188,953	16,228
US	5,599,940	0	0	0
Total	7,857,678	205,019	654,557	16,228

Table 2: Distribution on languages among the MAREC documents.

CLEF-IP corpus

The CLEF-IP track is part of the CLEF evaluation campaigns. It was launched in 2009 to investigate information retrieval (IR) techniques within the patent domain. There were two tasks in CLEF-IP 2010 campaign: Prior Art Candidates search and Classification. The first task consists in finding the prior art patent document for a given patent application. The latter consists of classifying a given patent document according to the IPC system up to the subclass level. The corpus provided for the tasks contains the bibliographic data, abstract, description, and claims. However, not all documents have content in all the fields. In MOLTO, we are interested in abstracts and claims. The text of the abstracts and claims is divided into several fragments, which are well marked. We are especially interested in those documents having the text fragments well aligned for the three languages.

The complete corpus has 2,680,604 patent documents, 822,144 of which are granted patents. 510,183 out of these have the claims translated into the three languages. There are 119,337 documents with IPC code A61P, the one we chose to represent the biomedical domain. Within this group, 21,150 granted patents have claims in the three languages which are aligned and will be used to build the parallel corpus. Table 3 gives a detailed numerical description of claims among the different IPC classes. Given that a patent can be classified into several classes, the sum up of the column may not be equal to the total number given in the last row. The first column corresponds to the total number of documents; columns 2 to 5 give the number of documents containing the claims in English (EN), German (DE), French (FR) or translated into the three languages (3Lang); columns 6 to 9 show the same figures for the fragments of the claims text. Note also that a patent

document could contain no translation of the claims, or the translation up to the three languages.

Class	Documents All	Documents EN	Documents FR	Documents DE	Documents 3Lang	Fragments EN	Fragments FR	Fragments DE	Fragments 3Lang
Total	2,680,603	1,210,390	719,265	857,188	510,183	21,847,114	12,158,335	13,816,441	7,241,843
A	531,813	223,467	141,916	162,440	103,703	4,487,951	2,870,826	3,132,035	1,691,592
B	735,095	334,434	212,295	262,628	156,567	5,319,142	3,197,071	3,795,064	2,078,287
C	698,581	302,126	177,636	209,969	134,718	5,965,095	3,620,469	4,015,630	2,239,391
D	84,426	37,130	24,187	31,079	18,824	590,747	391,811	485,444	268,590
E	110,618	45,577	31,960	42,587	23,566	636,444	427,473	565,376	287,069
F	319,485	142,040	93,133	115,800	67,109	2,034,743	1,244,894	1,498,487	788,151
G	664,426	311,855	167,615	186,700	108,961	6,551,588	3,205,516	3,445,392	1,714,031
H	590,089	282,814	152,337	171,326	97,303	5,688,167	2,689,959	2,911,963	1,395,565
A61P	*119,894*	*45,709*	*29,671*	*32,399*	*21,150*	*1,358,423*	*935,437*	*997,421*	*452,873*

Table 3: Number of documents having claims in English, German and/or French, and number of tokens in the claims for the same languages.

Table 4 and Table 5 give the numerical description of the abstracts. The former corresponds to the distribution of documents. Columns 1 to 3 give the number of documents containing the abstracts in English (EN), German (DE), French (FR). As can be seen in column 4 (3Lang), there are no documents with trilingual abstracts. Hence, we have also count the documents having bilingual abstracts, shown in columns 5 to 7. Table 5 gives the same figures for the fragments of the abstract text.

Class	Documents EN	Documents FR	Documents DE	Documents 3Lang	Documents EN-FR	Documents EN-DE	Documents FR-DE
Total	1,004,432	86,463	294,779	0	32,520	153,175	0
A	166,545	16,520	46,517	0	6,361	23,721	0
B	272,170	25,735	99,763	0	10,133	53,352	0
C	232,402	16,096	64,641	0	4,865	28,333	0
D	28,853	2,288	12,541	0	831	5,901	0
E	41,583	5,971	21,264	0	2,498	12,850	0
F	119,777	13,620	47,400	0	5,224	25,527	0
G	263,375	17,371	47,255	0	6,108	23,321	0
H	246,510	17,857	49,311	0	6,322	25,040	0
A61P	*29,520*	*2,401*	*5,734*	*0*	*600*	*2,129*	*0*

Table 4: Number of documents having abstracts in English, German and/or French.

According to these figures, one cannot build a trilingual corpus for translating abstracts, but the French-German translation can still be achieved by pivoting through English. Besides, the corpus of claims which is trilingual can be also used in this context.

Class	Fragments						
	EN	**FR**	**DE**	**3Lang**	**EN-FR**	**EN-DE**	**FR-DE**
Total	1,093,171	151,700	363,908	0	32,520	15,3175	0
A	185,423	28,447	57,957	0	6,361	23,721	0
B	292,385	44,444	122,256	0	10,133	53,352	0
C	266,439	29,901	83,816	0	4,865	28,333	0
D	31,762	4,065	16,286	0	831	5,901	0
E	44,382	10,195	25,623	0	2,498	12,850	0
F	127,164	23,702	57,917	0	5,224	25,527	0
G	284,898	31,095	58,004	0	6,108	23,321	0
H	264,299	31,936	60,163	0	6,322	25,040	0
A61P	*35,756*	*4,455*	*7,530*	*0*	*600*	*2,129*	*0*

Table 5: Number of fragments in the abstracts written in English, German and/or French.

3.2 Biomedical corpus

The first domain of application of the translation systems we chose in MOLTO includes biomedical and pharmaceutical patents. According to the IPC, we select patents with IPC code A61P, corresponding to the subclass "Specific therapeutic activity of chemical compounds or medical preparations".

As seen in the previous section, there are 21,150 granted patent documents with IPC code A61P with claims in English, French and German. There are no documents with abstracts in the three languages. Therefore, the corpus is build up with claims. Even though a patent has its claims in the three languages, it does not necessarily mean that those claims are aligned and can be used to build the corpus.

One can see in Figure 3 the structure of claims in the xml document. A claim is, in general, a long sentence slitted in fragments marked with <claim-text> tags, probably with nested elements. We search in every patent with trilingual claims, and count the number of *claim-text* elements. Whenever its number is the same for the three languages we assume that the claim is aligned and we add the aligned fragments to the corpus. So, our minimum aligned unit is shorter than a claim and, consequently, shorter than a sentence.

Even though fragments are shorter than a sentence, they may have a large number of words. For an appropriate use of the standard SMT software (`GIZA++` [ON03] and `Moses` [KHM+07]), the final corpus contains only those fragments with less than 100 tokens and with a ratio between the lengths of the source and target sentence less than 9. This methodology leads to 281,283 aligned parallel fragments as it can be seen in Table 6.

Beside, each of the fragments is cleaned in order to achieve an homogeneous corpus. Tags such as <sub> or
 are removed, chemistry formulae and images with the appropriate tag are substituted by ***IMAGE*** and extra spaces are removed for example.

The final parallel corpus is splitted in three parts. The largest part corresponds to the training corpus and has a total of 279,282 fragments and around 8 million tokens depending on the language (see Table 6). For every language, its side of the parallel corpus is used as a monolingual corpus to estimate the language model in the translation process. Two

SET	Fragments	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 6: Statistics for the patents parallel corpus on the biomedical domain in English (EN), German (DE) and French (FR).

smaller sets have been selected for development and test purposes, keeping 993 fragments for development and 1008 for test.

Notice that Table 6 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, as a collateral effect of the domain, some basic linguistic processors have also been built.

4 Out-of-domain corpus

Besides the biomedical patent corpus parallel in the three languages, we have compiled a large general-purpose bilingual corpus for training the basic SMT system. The corpus is a subset of the data sets included in the Sixth Workshop on Statistical Machine Translation (WMT11[11]) including European Parliament Proceedings (*Europarl*6) and Newspaper articles (*News*) which are available in English, French and German. The United Nations Proceedings have not been used as there is not the corresponding German translation. So, the final out-of-domain corpus contains mainly speeches from the Parliament Proceedings and to a lesser extent news; both of them domains with a very different grammatical structure and vocabulary to the biomedical patents domain.

Corpus	Fragments	EN tok	DE tok	FR tok
Europarl6 DE-EN	1,739,154	43,356,796	40,312,289	–
Europarl6 FR-EN	1,825,077	45,682,922	–	47,667,366
News DE-EN	136,227	2,909,872	3,006,634	–
News FR-EN	115,562	2,521,334	–	2,897,193
Total DE-EN	1,875,381	46,266,668	43,318,923	–
Total FR-EN	1,940,639	48,204,256	–	50,564,559

Table 7: Statistics for the out-of-domain aligned corpus in English (EN), German (DE) and French (FR).

Table 7 shows the number of aligned fragments in the aforementioned corpora. Notice

[11]http://www.statmt.org/wmt11/

that the number of aligned texts differs for different language pairs, being only available for German-English and French-English translation. However, German-French translation can be achieved by pivoting via English. The whole corpus is close to 2 million aligned fragments, almost an order of magnitude larger than the in-domain corpus. The analysis of the translation performance with both kinds of corpora will allow to study the trade-off between domain and size in the quality of translation. As in the case of the in-domain corpus, we use the corresponding size of the parallel corpus as monolingual corpus to build the language model.

5 Summary

This document reports on the collection of corpora gathered for training the systems developed within WP5. The goal of the workpackage is to develop several approaches to hybridisation between rule-base and statistical machine translation techniques. We have two base systems: the GF, being further developed within WP3, and a state-of-the-art phrase-based statistical machine translation system.

The translation of patents is a case of study where research is focused on simultaneously obtaining a large coverage without loosing quality in the translation. We undertake the translation of patents into three languages (English, French and German) within the biomedical and pharmaceutical domains, a natural scenario to test the techniques developed in this workpackage.

Both translation systems in MOLTO, and in particular the statistical one, require of parallel corpora to train the translation models and build specialised lexicons, and monolingual corpora for the language models.

On the one hand, we have gathered a general purpose out-of-domain corpus from several public sources, such as the European Parliament corpora and the United Nations Proceedings. On the other hand, we have build up a specialised in-domain corpus containing patent documents pertaining to the "therapeutic activity of chemical compounds or medical preparations" class (IPC code A61P).

Due to the lack of the EPO corpus, up to now we are using a research license of the CLEF-IP data. Since we have build the corpus from the subset of patents pertaining to EPO, we expect to keep the described corpus during all MOLTO lifetime. In any case, the present deliverable will be updated as soon we obtain the final corpus or an approval for the current one.

References

[KHM+07] Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Transla-

tion. In *Annual Meeting of the Association for Computation Linguistics (ACL), Demonstration Session*, pages 177–180, Jun 2007.

[ON03] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[Ran11] Aarne Ranta. *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford, 2011. ISBN-10: 1-57586-626-9 (Paper), 1-57586-627-7 (Cloth).

[TÏ1] Wolfgang Täger. The Sentece-Aligned European Patent Corpus. *Proceedings of the 15th Annual Conference of the European Association for Machine Translation (EAMT 2011)*, 2011.