# Translating the Unitary Patent I: ‘Laminated Jealous Glass’

**Kluwer Patent Blog**
December 18, 2014

Kluwer Patent blogger

*Please refer to this post as: Kluwer Patent blogger, ‘Translating the Unitary Patent I: ‘Laminated Jealous Glass’’, Kluwer Patent Blog, December 18 2014, http://patentblog.kluweriplaw.com/2014/12/18/translating-the-unitary-patent-i-laminated-jealous-glass/*

---

For Europeans who don’t speak English, German or French, the three official Unitary Patent (UP) languages, the future UP system will bring about an even more radical change than for those that do. Over the years, millions of patents from companies all over the world will have been held valid in their territory, although these patents would only be available in one of those three official UP languages. With Patent Translate, the machine translation system developed jointly by the European Patent Organisation (EPO) and Google, anyone will henceforth be able to read a patent description in his or her mother tongue.


Patent Translate has a statistical approach. ‘The system translates by comparing sentence by sentence from a source document to millions of patent documents which have previously been translated by human translators for the purposes of preparing patent specifications. The system is equipped with a “learning” facility based on official patent documents collected by the EPO in cooperation with patent offices in Member States (…)’, as Deloitte explains in its 2012 report ‘Analysis of prospective economic effects related to the implementation of the system of unitary patent protection in Poland’.

At first glance, all is seemingly going well. As can be read on the EPO’s website: ‘By the end of 2014, machine translation of patents will be available for the languages of the 38 Member States of the European Patent Organisation, including the European Union’s 27 Member States.’

But what exactly does ‘available’ mean? What about the quality of the translations? In an earlier report on this blog, we saw that, for the Czech Republic at least, the language issue is the most serious problem of the UP package. It is ‘necessary to work on improving the quality of the machine translations into Czech to provide our users of patent information with wording understandable in their own language’, said Josef Kratochvil, president of the Czech Intellectual Property Office (IPO). Patent Translate ‘provides correct Czech expressions but, in such a way (sequence, etc.) that it is sometimes not possible to understand even the field of technique, let alone the claims.’

Several other UP Member States know all too well what Kratochvil is talking about. During a meeting on the issue on 23 and 24 October 2014 in Prague, Finnish representatives made clear that the English-Finnish machine translation provided by the EPO is at the moment ‘useless’. In addition, Mr. Csaba Baticz, deputy head of the Legal and International Department of the IPO of Hungary, told Kluwer IP Law: ‘The unanimous judgment of our examiners is that the current machine translations are at least of a ‘rather bad’ quality, between 3 and 4 on a scale of 10.’

Mr. Jorma Hanski, director of the Patents and Innovations Line of the Finnish IPO, gives the translations a ‘2 to 3 on a scale of 10’. ‘A short test with EPO’s English-Finnish machine translation indicated that less than 10 percent of the translated sentences were correct and less than half were comprehensible for a native Finnish speaker.’

In a recent article, Hana Churackova, head of the electronic services of the Patent Information Department of the Czech IPO, explained why the phrase-based statistical machine translations of Patent Translate don’t work too well for her mother tongue:

‘For development of the translator and the switch from one language to another language, Google used the so-called language pairs provided by the European Patent Office. For most languages, we can assume a direct proportion to the effect that the more language pairs, the more perfect translation.

(…) Why is it so difficult to create a machine translator for making translations from English into Czech? I certainly do not think it is just because of lack of the respective language pairs. English belongs among the so-called isolating language types. Suffixes are not used for declension and conjugation, words are essentially invariable. Sentence structure is formed by word order (…).

The Czech language, unlike the English language, belongs among the so-called flexible language types and has a number of irregularities and exceptions. Flexible languages typically accumulate meanings in a single morpheme which is able to express, at the same time, multiple functions, such as case, number and gender. Let’s have a look at the adjective “good“, for example. In English, a woman, a man, a child or a machine is always “good“, whether in the singular or in the plural, in the nominative, genitive or dative. It is different in Czech. A woman is “dobrá“, a man is “dobrý“, a child is “dobré“. And that’s just the singular nominative. (…) Another problem is posed by nouns and their case endings (there are seven cases in the Czech language), in which functions are accumulated again (…). While declining, there often occurs a change of stem vowels. Not only nouns but also adjectives, pronouns and numerals are declined, thus having various forms. Word order is free which, (…) causes the language to be highly sensitive to functional sentence perspective. And this is just an outline of some problems (…).’

In Patent Translate, Churackova found several rather amusing translations: ‘a battery which promotes safety by reducing hate…’, ‘laminated jealous glass’, ‘a fine Bohr envy plate’, ‘self-assessment wizard’. She doesn’t know how the linguistic issues can be solved, but proposes, as a start, to remove terms expressing human emotions, such as hate, jealousy, anger, envy, or expressions used for fairy-tale characters from the dictionary of synonyms.

Although 60.000 Hungarian and English document pairs have been processed in Patent Translate, problems as well occur with translations from/to Hungarian and English. ‘Its morphological processes yield a huge number of different word forms. This, combined with free word order of main grammatical constituents and systematically different word order, results in poor performance of traditional phrase-based Statistical Machine Translation systems’, according to Csaba Baticz.

The situation was aggravated because 35.000 document pairs were available only on a paper carrier and were first digitalized through Optical Character Recognition (OCR), which led to additional mistakes being built into the translations. The word ‘on’, for instance, was sometimes recognized erroneously as ‘oil’ and introduced as a new technical feature.

Despite all this, Baticz emphasized ‘Patent Translate is a definite improvement in comparison with Google Translate’. Furthermore, he explained, the EPO presented a project on data acquisition under Quality at Source at the October meeting in Prague:

‘Under the scope of this project, a front file delivery of patent data will be established in an EPO defined form based on the concept of Quality at Source (Q@S). When the front file delivery is well established for an NPO, fulfilling all the EPO quality criteria, then the missing back file patent data from 1973 to date will be collected in digital format covering bibliographic, image and full-text data (full-text format when the quality of the original document allows it). The outcome of this project will benefit all parties: the EPO, the NPOs and the public. The project will provide additional patent corpora, which could be used to further improve the quality of the Patent Translate service.’

According to Hana Churackova, ‘both the EPO and Google are aware of all pitfalls associated with translations and intend to focus their effort on improving quality, so results of machine translations come closer to human translations’.

Jorma Hanski of Finland, however, is less optimistic: ‘It seems that statistical machine translation engines (such as Google Translate used by the EPO) work poorly with the Finnish language. Better results may be achieved with rule-based or hybrid machine translation technologies.  It is not clear whether or not Google is interested in adopting such technologies to improve the machine translations from/to Finnish.’

*Next week, in a second blogpost on machine translations, we’ll focus on the question what consequences the system of machine translations will have for doing business.*